Univerzita Karlova v Praze

Filozofická fakulta

Parametrická syntéza a percepční ověření českých vibrant

Parametric synthesis and perceptual verification of Czech trills

Bakalářská práce

Martina Koppová

Praha, 2011

Vedoucí práce **Mgr. Radek Skarnitzl, Ph.D.**

Poděkování

Děkuji vedoucímu bakalářské práce Mgr. Radku Skarnitzlovi, Ph.D. za metodické vedení práce. Dále děkuji svému manželovi Matyáši Koppovi za technickou pomoc se sestavením online verze percepčního testu.

Prohlašuji, že jsem bakalářskou práci vypracovala samostatně, že jsem řádně citovala všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze dne 19.8.2011 *podpis*

Abstrakt

Pro parametrickou syntézu pomocí high-level parametrů v systému HLSyn existuje parametrický popis hlásek americké angličtiny, nad níž byl systém autory Stevensem a Bickleym vytvořen jako typ syntézy založený na formantové syntéze Denise Klatta a kombinující vlastnosti formantové a artikulační syntézy.

Cílem této práce bylo sestavit parametrický popis českých vibrant, tedy /r/ a /ř/, provést syntézu těchto hlásek a otestovat jejich přirozenost percepčním testem.

Percepční test ukázal, že jednoduchá vibranta /r/ lze syntetizovat bez větších problémů na základě úprav parametrů znělé alveolární explozivy, aniž by byly tyto úpravy vnímány posluchači jako nepřirozené. Oproti tomu frikativní vibranta /ř/ je hláska natolik komplikovaná, že i přes snahu věrně okopírovat její průběh podle reálných dat ji posluchači hodnotili jako nepřirozeně znějící.

Abstract

There is a parametric description of American English sounds suitable for parametric synthesis using high-level parametres in the HLSyn system, upon which the Stevens and Bickley system was created as the type of synthesis based on Denis Klatt´s formant synthesis and combining the aspects of formant and articulatory synthesis.

The aim of this work was to create a parametric description of Czech vibrants, i. e. /r/ and /ř/, synthesize these sounds and examine whether they are considered natural or not in a perception test.

The perception test has shown that it is possible to synthesize the simple vibrant /r/ without any problems; the synthesis was based on modification of sounded alveolar explosive parametres and the modification was perceived as natural. However, the fricative vibrant /ř/ proved to be too complicated to be synthesized and even though we tried to follow its development according to real data it was still perceived as unnatural.

Klíčová slova: parametrická syntéza, HLSyn, vibranty

Key words: parametric speech synthesis, HLSyn, vibrants

OBSAH:

1 Úvod

V naší práci se budeme zabývat parametrickou syntézou českých vibrant. Systém HLSyn, ve kterém budeme pracovat, je totiž vyvinut pro angličtinu a vibranty v něm nejsou teoreticky popsány. Pro syntézu likvid (l, aproximantní r) se využívá pouze změn v formantech, což pro vibrantní likvidy není možné.

Teoretický úvod je rozdělen na dvě části – první se zabývá metodami syntézy řeči s důrazem na parametrickou syntézu, druhá uvádí přehled vlastností r-ových hlásek (rhotics).

Experimentální část práce sestává z popisu variování parametrů při syntéze vibrant /r/, znělého /ř/ a neznělého /ř/ a percepčního testu – jeho popisu a výsledků, kterým byla přirozenost percepce jednotlivých variant ověřena u rodilých mluvčích češtiny.

1.1 Metody syntézy řeči

V dnešní době širokého využití vyspělých IT technologií místo lidského potenciálu je syntéza řeči jednou z oblastí zájmu, v níž je stále možno dosáhnout výrazného zlepšení. Protože hlasová komunikace je jedním ze základních pilířů lidského počínání, spočívá snaha pozvednout komunikaci se stroji na úroveň mezilidské komunikace především právě v možnosti domluvit se s nimi pouze hlasově, bez nutnosti mechanického ovládání (kam patří i použití klávesnice). K tomu je nezbytně nutná zejména co nejlepší úroveň rozpoznání toho, co člověk říká, tedy jakási „strojová percepce“, ovšem i druhý aspekt, „strojová produkce“, je k úplné imitaci mezilidské komunikace potřeba. Cílem je imitace lidské řeči tak, aby byla od přirozené řeči k nerozeznání. Těžko posoudit, zda je tato meta kompletně dosažitelná, jisté je, že současné metody dokáží syntetizovat řeč tak, že je srozumitelná a pro praktické účely použitelná, avšak stále působí značně nepřirozeně.

Syntetizéry je možno dělit do skupin podle dvou hledisek – zaprvé podle využité techniky na

a) mechanické,
b) elektronické,
c) digitální,

toto rozdělení bude popsáno v kapitole Historický přehled (1.1.1), nebo podle principu syntézy na

a) konkatenační,
b) parametrické artikulační,
c) parametrické formantové,

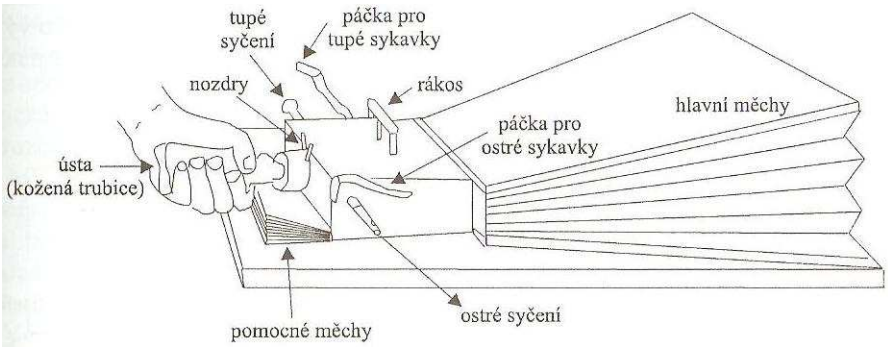
což bude detailněji rozebráno v kapitole Dnešní metody (1.1.2).

Není-li uvedeno jinak, je zdrojem této kapitoly Psutka (2006).

1.1.1 Historický přehled

1.1.1.1 Mechanické syntetizéry

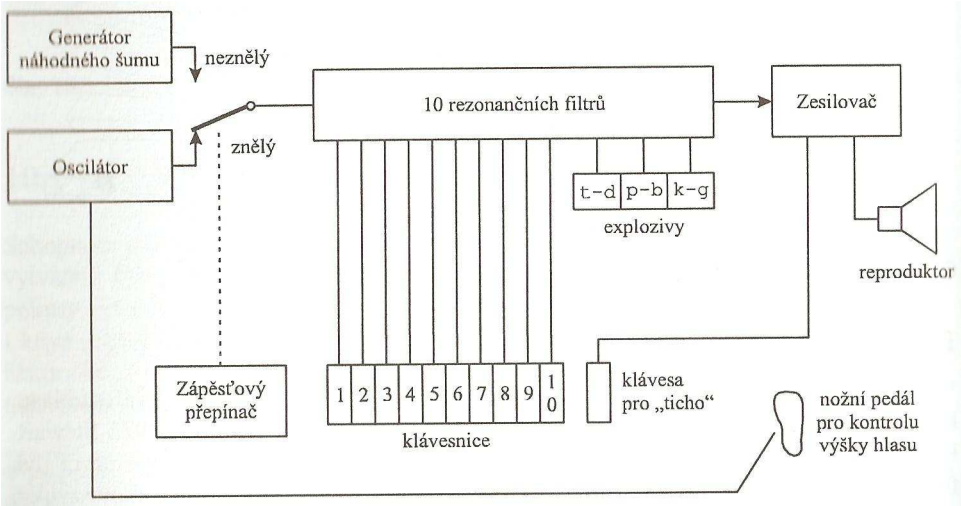
Prvním pokusem o syntetizér lidské řeči byl přístroj Ch. Kratzensteina z roku 1779, který dokázal syntetizovat vokály pomocí rákosového plátku (podobného jako u hudebních nástrojů) a série rezonátorů. Dalším historickým pokusem, který už „syntetizoval“ i některé konsonanty, byl přístroj W. von Kempelena, který byl sestaven z kožené trubky imitující vokální trakt a opět rákosové píšťalky, která zastupovala hlasivky (obr. 1). Pomocí změn tvaru kožené trubky dokázal měnit vokály. Tento přístroj vylepšil Ch. Wheatstone, jeho stroj se ovládal systémem páček a bylo na něm prý možno již syntetizovat jednoduchá slova. Výsledky syntézy pomocí strojů W. von Kempelena Ch. Wheatstonea ukázaly na skutečnost, že na modulaci řečového signálu se vysokou měrou podílí tvar vokálního traktu, nejen hlasivky.



Obr.1 Přístroj podle W. von Kempelena (převzato z Psutka, 2006)

1.1.1.2 Elektronické syntetizéry

V roce 1939 byl v Bellových laboratořích vyvinut první elektronický syntetizér souvislé řeči, „Voder“ (Voice Operating Demonstrator). Systém fungoval na principu 10 rezonátorů
ovládaných potenciometry (obr. 2).



Obr.2 Schéma systému Voder (převzato z Psutka, 2006)

Jeho obsluha byla velice náročná, ovšem bylo možno produkovat již srozumitelnou souvislou řeč, i když dnešní úrovni byla kvalita ještě velmi vzdálena. Tento přístroj je jako první možné nazývat skutečným syntetizérem řeči.

V roce 1953 byl představen první formantový syntetizér PAT (Parametric Artificial Talker) Britem W. Lawrencem, a brzy na to kaskádový formantový syntetizér OVE (Orator Verbes Electris) Gunara Fanta. Ten byl založen na Fantově teorii zdroje a filtru (Fant, 1960). Tato teorie říká, že je možné nezávisle modelovat zdroj buzení a filtr popisující vokální trakt. (více viz Formantová syntéza 1.1.2.2.1)

Počátkem padesátých let představil Franklin Cooper systém Pattern Playback, předchůdce dnešní konkatenační syntézy.

První artikulační syntetizér se objevil v roce 1958 pod názvem DAVO, pro svou obecnou složitost však výhradně artikulační syntetizéry nenašly příliš pokračovatelů.

1.1.1.3 Digitální syntetizéry

Počátkem 60. let se s rozvojem komputační vědy otevřely nové obzory i pro oblast syntézy řeči – místo hardwarové realizace modelů bylo možné řeč modelovat číslicově. Modely OVE a PAT byly přepracovány do digitální podoby a bylo stvořeno mnoho dalších nových modelů.

Zatímco do této chvíle se syntéza řeči zabývala pouze modelováním řeči obecně, od 60. let se začaly vyvíjet modely syntézy řeči z textu (text-to-speech, TTS), které ke své realizaci potřebují také jazykový model kvůli rozdílům mezi ortografií a skutečnou výslovností. Přepis může být stanoven dvěma způsoby, a to buď na základě fonetického slovníku (hodí se pro jazyky s historickým pravopisem a analytického typu – např. angličtinu), kde je ke každému slovu uvedena jeho výslovnost, či na základě souboru systematických pravidel (hodí se naopak pro jazyky s fonologickým pravopisem a flektivního typu – např. pro češtinu).

Roku 1976 byl představen první prakticky využitelný syntetizér – používal se pro předčítání tištěného textu nevidomým.

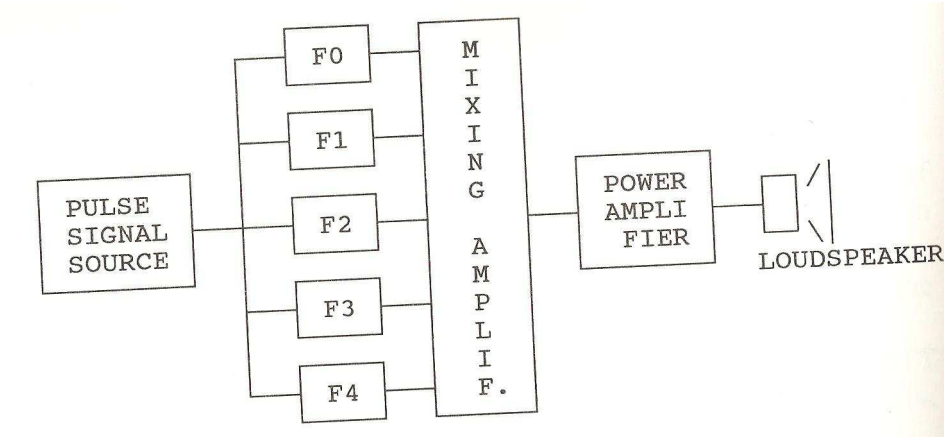
V 70. letech přišel Denis Klatt z MIT s formantovým syntetizérem MITalk, v roce 1982 s jeho vylepšenou verzí pod názvem Klattalk. Jeho technologie se stala základem mnoha dalších systémů, včetně systému HLSyn, který využíváme v naší práci. (O systémech Klattalk a HLSyn detailněji kapitoly Klattovská syntéza 1.1.2.2.2.2 a HL technologie 1.1.2.2.3). Dalším významným systémem založeným na Klattalku byl komerční systém DECTalk vytvořený společností Digital Equipement Corporation roku 1983, pravděpodobně nejpoužívanější TTS systém 20. století.

Stejně tak konkatenační modely získaly s rozvojem počítačů lepší prostor pro realizaci. V roce 1985 představují Charpentier a Moulines z France Telecom techniku PSOLA, metodu pro modifikaci prozodických charakteristik řetězených jednotek.

Od 90. let se syntéza orientuje především na korpusově orientovanou syntézu řeči – korpusy jednotek pro řetězení se vytvářejí automaticky z velkého objemu dat. K tomu se využívá metod rozpoznávání řeči, především skrytých Markovových modelů (Hidden Markov Models, HMM), nebo neuronových sítí (NN).

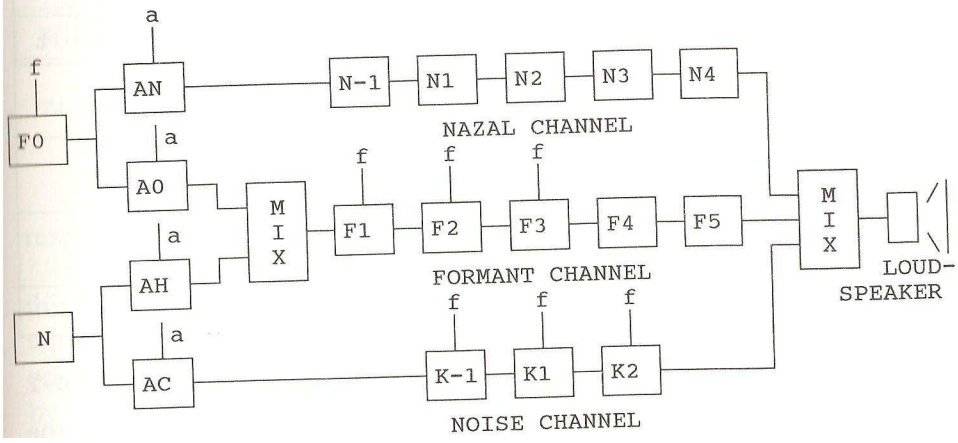
1.1.1.4 Historie české syntézy (Ptáček, 1996)

První pokusy o syntézu českých samohlásek provedl ve 20. letech 20. století František Kaňka. První skutečný syntetizér českých stacionárních zvuků sestavil v 1964 Přemysl Janota (FÚ FF UK). Jednalo se o jednoduchý elektronický formantový syntetizér, který uměl syntetizovat pouze vokalické pole. Byl využíván především pro výzkum hledisek percepčních testů (obr. 3).



Obr. 3 Schéma formantového syntetizéru P. Janoty (zleva zdroj pulsového signálu, F0-F5 rezonátory simulující základní frekvenci a jednotlivé formanty, sloučení signálu z jednotlivých rezonátorů, zesilovač, reproduktor) (převzato z Ptáček, 1996)

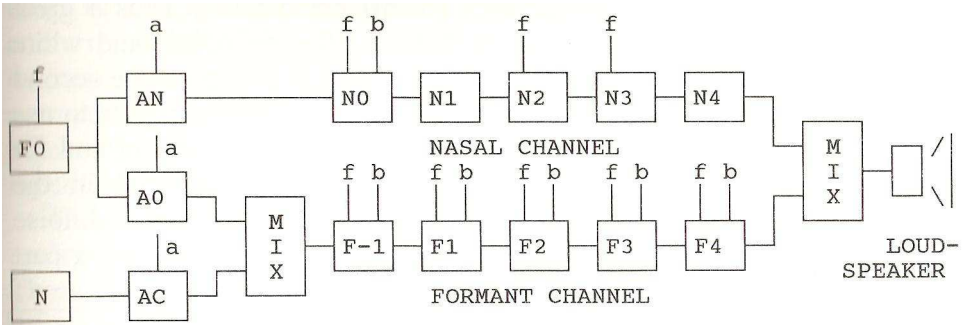
V roce 1968 představili B. Borovičková a V. Maláč ve spolupráci s Výzkumným ústavem TESLA první syntetizér souvislé češtiny OVED1. Stroj byl založen na principech modelu OVE II Gunara Fanta, , poslední písmeno zkratky značí digitální technologii (obr. 4).



Obr. 4 Schéma syntetizéru OVED1 (f – nastavení frekvence rezonátoru, a – nastavení útlumu F0 – zdroj tónu, N – zdroj šumu, A0 – amplituda tónu pro formantovou větev, AH – amplituda šumu pro formantovou větev, AN – amplituda tónu pro nazální větev, AC – amplituda šumu pro šumovou větev, N-1 nazální antirezonátor, N1-N4 nazální rezonátory, F1-F5 formantové rezonátory, K-1 šumový antidetonátor, K1-2 šumové rezonátory, MIX – sloučení frekvencí z jednotlivých větví, reproduktor) (převzato z Ptáček, 1996)

Byl založen na kaskádovém zapojení Helmholtzových rezonátorů/antirozonátorů ve třech paralelních větvích. Jedna větev syntetizovala formantovou strukturu (F1-F5), přičemž první tři formanty se daly průběžně nastavovat, zatímco zbylé dva byly stabilně nastaveny na pevnou hodnotu. Mezi nejzásadnější počiny s tímto systémem patřila první simulace všech českých hlásky v intervokalickém okolí. Výsledky byly otištěny v Maláč, V. (1967) Speech synthesis. Výzkumná zpráva Výzkumného ústavu pro sdělovací techniky (dnes nedostupné) a přetištěny v Borovičková, Maláč (1967). Tyto výsledky uvádíme v části věnované vibrantám (Vibranty 1.2.3) a v experimentální části práce. .

V roce 1975 byla představena vylepšená verze OVEDu, nazvaná HO2. Vylepšení spočívalo ve dvou aspektech – modelování vokálního traktu jako polouzavřeného tubusu spíše než série Helmholtzových rezonátorů a možnosti využití pouze dvou větví místo tří, protože se spojila větev formantová a šumová (obr. 5).



Obr. 5 Schéma syntetizéru OH2, popis komponent analogický systému OVED1, b – nastavení šířky pásma (převzato z Ptáček, 1996)

Dalšími modely byly HO3 (1973) a HO4 (1977), které se lišily využitou technikou ovládání parametrů – byly již ovládány počítačově.

První konkatenační syntetizér češtiny sestavili v roce 1972 M. Ptáček, V. Maláč a P. Dvořák.
V 90. letech byl ve spolupráci Fonetického ústavu a Ústavu radiotechniky a elektroniky AV ČR vyvíjen parametrický konkatenační systém TTS využívající manuálně připravovaný inventář difonů. V současné době je na Ústavu fotoniky a elektroniky AV ČR vyvíjen systém s názvem EPOS.
Od konce 90. let se syntézou řeči začala zabývat Katedra kybernetiky Západočeské univerzity v Plzni (J. Matoušek, J. Psutka), kde byl vytvořen systém ARTIC (Artificial Talker in Czech), první český korpusově založený konkatenační systém, využívající techniky HMM k automatickému rozpoznávání řečových jednotek do inventáře. Tento systém je založen na trifonech.

1.1.2 Dnešní metody

Dnes lze v oblasti syntézy řeči nalézt dva zcela odlišné postupy – konkatenační syntézu a parametrickou syntézu, která se dělí na formantovou a artikulační. Každá z metod má své výhody a nevýhody.

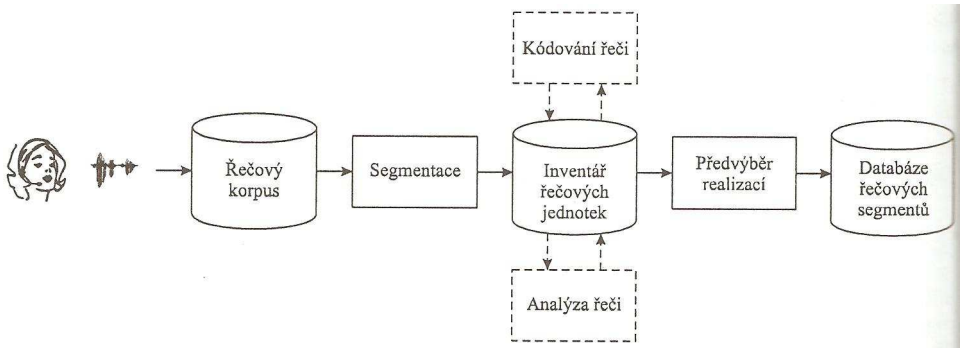
1.1.2.1 Konkatenační syntéza

Konkatenační syntéza je dnes nejvíce používaným přístupem k syntéze řeči. Je založena na předpokladu, že jednotlivé zvuky, z nichž se řeč skládá, lze reprezentovat pomocí konečného počtu řečových jednotek. Tyto řečové jednotky jsou pak uloženy v inventáři, jímž je většinou označkovaný korpus.

1.1.2.1.1 Princip konkatenační syntézy

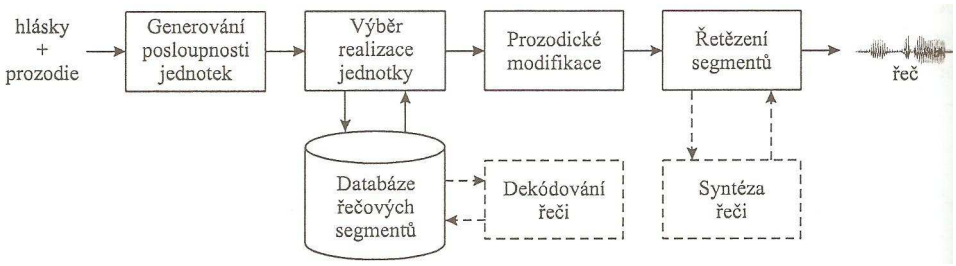
Před samotnou syntézou je zapotřebí několika přípravných kroků. Zaprvé, zvolit typ řečové jednotky (slabiky/fonémy/difony atd. viz Typ řečové jednotky 1.1.2.1.2), poté vybrat fráze, jež bude korpus obsahovat (je nutné, aby každý prvek inventáře byl zastoupen minimálně jednou, lépe však vícekrát). Následuje segmentace frází, ruční či

automatická. K automatické segmentaci více níže (Metody založené na skrytých Markovových modelech 1.1.2.1.3). Před syntézou je také dobré vybrat zástupce prvků (pokud se v korpusu nacházejí vícekrát) a uložit je do databáze. Pokud chceme parametry segmentů dále upravovat (parametrická konkatenační syntéza), je nutné provést analýzu řeči, při které jsou jednotlivé prvky převedeny na soubor vektorů jednotlivých parametrů (obr. 6).



Obr. 6 Princip vytváření inventáře řečových jednotek (přerušované rámečky značí fakultativní součásti) (převzato z Psutka, 2006)

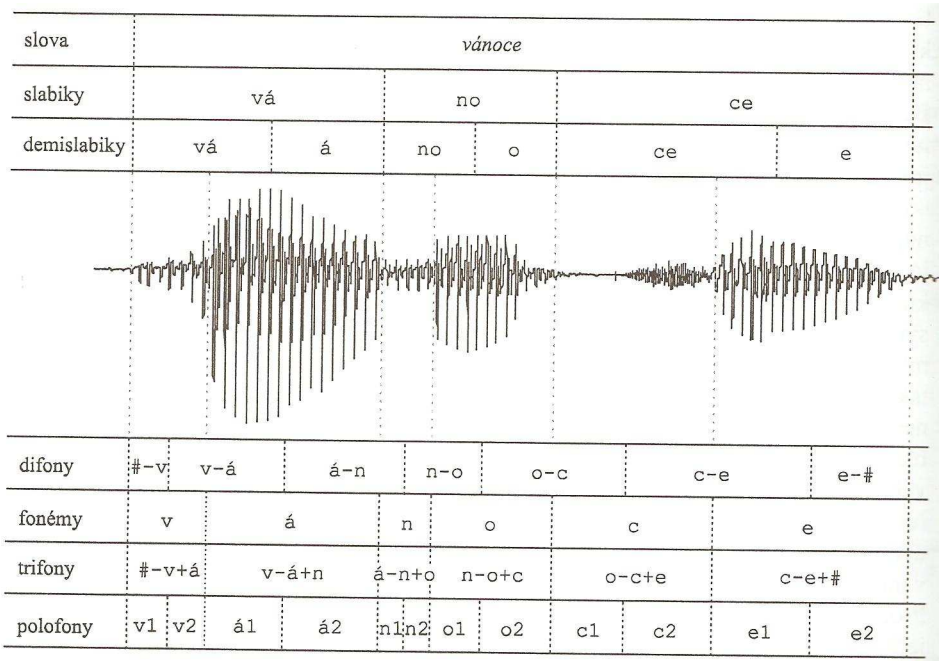
Když je databáze připravena, nastává samotná fáze syntézy řeči. Vstupem je textově zaznamenaná budoucí promluva, neboli posloupnost hlásek, případně doplněná prozodickými značkami. Systém pak každé hlásce přiřadí z databáze odpovídající prvek, pokud databáze obsahuje více prvků odpovídajících dané hlásce, vybere tu, která nejlépe odpovídá kontextu zamýšlené hlásky. Poté systém může provést prozodické modifikace, aby nedocházelo k prozodickým skokům mezi segmenty. Při samotném řetězení, které následuje, vyhlazuje spektrální přechody, aby nedocházelo ve spektrální oblasti ke skokům. Pokud je v inventáři velký výběr jednotek, vybírá je systém rovnou tak, aby na sebe prozodicky i spektrálně navazovaly a vyhlazování pak provádět nemusí (obr. 7).



Obr. 7 Princip konkatenační syntézy (přerušované rámečky značí fakultativní součásti) (převzato z Psutka, 2006)

1.1.2.1.2 Typ řečové jednotky

Při výběru řečové jednotky proti sobě stojí dva páry kritérií. První pár, požadavek na maximální pokrytí koartikulace a minimálně problémy s prozodickými a spektrálními nespojitostmi, mluví pro co největší jednotky, tedy v extrémním případě fráze až věty. Druhé dva požadavky, a to požadavek zobecnitelnosti na jakýkoliv syntetizovatelný text a zároveň únosně veliká databáze jednotek svědčí naopak pro co nejmenší jednotky, fonémy až polofony (obr. 8). Výběr typu také závisí na vlastnostech jazyka, pro češtinu jsou nejčastěji používané difony (systém EPOS), neboli segment začínající v polovině první hlásky a končící v polovině druhé hlásky. V systému PSOLA jsou užívanou jednotkou trifony, neboli segmenty zahrnující celou cílovou hlásku, polovinu hlásky předchozí a polovinu hlásky následující.



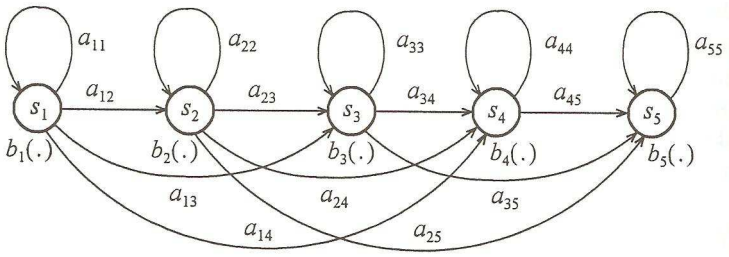
Obr. 8 Příklad typů řečových jednotek (převzato z Psutka, 2006)

1.1.2.1.3 Metody založené na skrytých Markovových modelech

Statistická technika skrytých Markovových modelů se původně využívala k rozpoznávání řeči, dnes se jí využívá i v syntéze, a to dvěma způsoby.

1.1.2.1.3.1 Princip HMM

Markovovské modely jsou založené na množině stavů a na množině pravděpodobností přechodu mezi těmito stavy. Pro modelování mluvené řeči se používají tzv. levo-pravé modely, které jsou vhodné pro modelování procesů probíhajících v čase. Model začíná ve stavu s nejnižším indexem a podle vývoje přechází buď do stavu s vyšším indexem, nebo setrvává ve stejném stavu. Průchod modelem je tedy graficky reprezentován zleva doprava (obr. 9).



Obr. 9 Příklad pětistavového modelu (převzato z Psutka, 1996)

Původně stavy reprezentovaly jednotlivá slova, ovšem ukázalo se, že využití menších jednotek systém značně zjednodušuje. Dnes se používají modely jednotlivých fonémů doplněné o model ticha.

1.1.2.1.3.2 Využití HMM pro tvorbu inventáře jednotek

Pro využití velkého korpusu dat, který zajistí širší spektrum možností syntézy, by bylo zapotřebí příliš mnoho manuální práce s přípravnou segmentací jednotek. Segmentace se tak provádí automaticky za využití moderních algoritmů rozpoznávání řeči. Tyto metody jsou dvě – pomocí skrytých Markovovských modelů (HMM) a pomocí techniky borcení časové osy (DTW), ta je založena na porovnání přirozené řeči se syntetizovanou řečí stejného obsahu, v níž hranice jednotek známe. HMM se vyznačuje konzistentní segmentací, ale průměrná chyba segmentace bývá větší. DTW je náchylná k vytváření velkých segmentačních chyb, což je pro syntézu závažnější. Více se tedy zatím používá metody HMM.

Rozpoznávání pomocí HMM zahrnuje dvě fáze – fázi trénovací, kdy si systém na základě označkovaného korpusu promluv vytvoří soubor vektorů, které reprezentují stavy, a pravděpodobností přechodů mezi těmito stavy; a fázi samotného rozpoznávání, kdy systém dostane neznámou promluvu a na základě svého modelu simuluje

nejpravděpodobnější průchod jednotlivými stavy, neboli fonémy, který odpovídá ručně nadefinovanému jazykovému modelu (např. slovník povolených slov).

Hranice fonémů jsou pak místa, kde model přechází z jednoho stavu do dalšího. Pro využití rozpoznávání pomocí HMM za účelem utvoření označkovaného korpusu pro následnou konkatenační syntézu založenou na jiné řečové jednotce než fonému (např. difonu), se do korpusu ukládají vždy sousední dvě hlásky a jejich hranice se stanovuje až při syntéze, buď jako polovina trvání, nebo tak, aby se minimalizovaly nespojitosti vzniklé během syntézy.

1.1.2.1.3.3 Využití HMM pro samotnou syntézu

Od roku 2002 je japonskými vědci (Tokuda, Oura, Zen) vytvářen systém syntézy přímo pomocí HMM. Princip je podobný jako u rozpoznávání – nejprve se korpusu promluv extrahují parametry a na jejich základě se vytvoří model pro každý foném a pravděpodobnosti přechodů mezi nimi. Při samotné syntéze jsou pak generovány parametry syntetizované řeči tak, aby pravděpodobnost průchodu natrénovanými modely byla maximalizována. Tato metoda byla sice vyvinuta pro japonštinu, ale protože je jazykově nezávislá, byly už provedeny první pokusy s její aplikací na češtinu. (Hanzlíček, 2010)

1.1.2.1.4 Výhody a nevýhody

Výhodou konkatenační metody je její jednoduchost a relativně vysoká efektivita, jelikož nevyžaduje nastavování parametrů ani znalosti procesu tvoření řeči. Také není problém s přirozeností barvy hlasu, jelikož ta vyplývá z reálných promluv konkrétního řečníka. Na druhou stranu, tato závislost na řečníkovi představuje značné omezení, protože změna barvy hlasu je sice technicky možná přes parametrizaci, ovšem existuje riziko, že se změní i další charakteristiky. Dalším nebezpečím konkatenační syntézy je riziko špatné kvality segmentu v korpusu. Tomu se dá předcházet jedině pořízením velikého korpusu a výběrem nejvhodnějšího kandidáta. Největším zdrojem potenciálních problémů jsou pak místa zřetězení, kde může dojít k prozodickým či spektrálním skokům. V neposlední řadě má tato metoda vysokou paměťovou náročnost, v případě parametrizace a parametrické úpravy i výpočetní náročnost.

Tato metoda je tedy relativně efektivní, a proto dnes široce využívaná, má však své hranice, které lze překročit jen za cenu výrazného zvýšení složitosti procesu a tím výpočetní náročnosti.

1.1.2.2 Parametrická syntéza

Pod parametrickou syntézu shrnujeme dva typy – syntézu artikulační a formantovou. Parametrická syntéza je obecně založena na počítačové modulaci akustické vlny. Akustickou vlnu je tedy nutno několika parametry popsat. Tyto parametry mohou být buď čistě akustické, založené pouze na fyzikálních vlastnostech zvuku (formantová syntéza) nebo na artikulačním nastavení mluvidel. Zatímco formantová syntéza je značně neintuitivní, ovšem technicky méně náročná, artikulační syntéza je bližší reálné produkci řeči, protože odráží skutečné nastavení vokálního traktu. Je pak ovšem potřeba převod hodnot nastavení vokálního traktu na matematické rovnice vyjadřující akustickou kvalitu. To je teoreticky možné, protože artikulační a akustická rovina spolu v mnoha ohledech souvisejí, ovšem natolik náročné, že je pak výpočetně a metodicky mnohem složitější než metoda formantová.

1.1.2.2.1 Artikulační syntéza

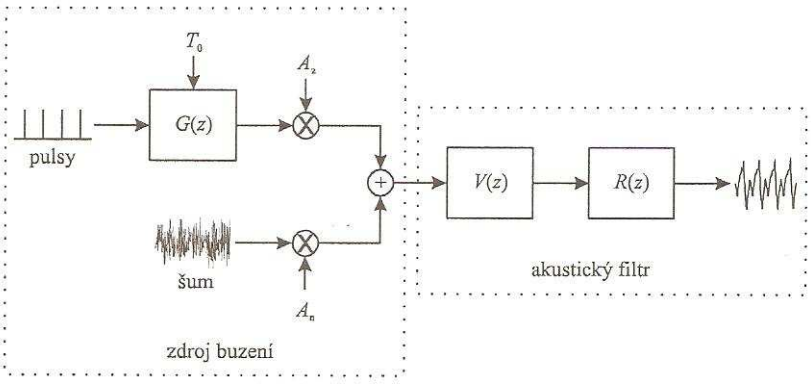
Artikulační syntéza si všímá reálných parametrů lidského mluvního ústrojí, například přesné vzdálenosti rtů, přesných pohybů hlasivek atd. Tento přístup je nejblíže napodobení reálného zvuku, všímá si skutečně předmětu procesu vytváření řeči, na rozdíl od konkatenační syntézy, která tento aspekt téměř neřeší, a formantové syntézy, která vychází z kvalit zvuku, tedy až po procesu jeho vytvoření. Největší překážkou je výpočetní náročnost (je potřeba nelineárních parciálních diferenciálních rovnic) a obtížné zjišťování některých artikulačních parametrů, k nimž je nutné použít rentgen nebo magnetickou rezonanci.

Ačkoli by se mohlo zdát, že metoda, která jde přímo k meritu věci a snaží se o co nejpřesnější napodobení, by se jednou mohla dopracovat nejlepších výsledků, není to nutným pravidlem. Často i metoda, která využívá jiného principu, dokáže imitovat skutečnost lépe – „vždyť i letadla umějí létat a nemají křídla“.

Principu artikulační syntézy a především blízkosti artikulační a akustické charakteristiky řeči však částečně využívá HLSyn metoda, o níž bude řeč níže (HL technologie 1.1.2.2.3)

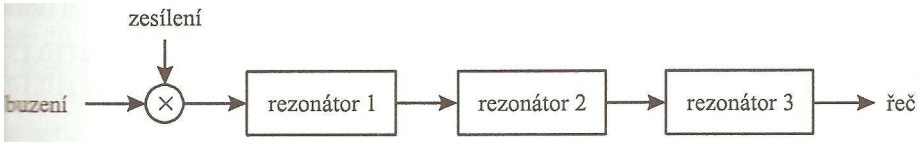
1.1.2.2.2 Formantová syntéza

Formantová syntéza je založena na teorii zdroje a filtru (Fant, 1960), neboli na modelování zdroje buzení a filtru imitujícího vokální trakt. Vychází z předpokladu, že každý zvuk lze matematicky popsat jeho spektrem jeho zdroje G(f) (kde f jsou hodnoty v Hz), což zastupuje přirozené zdroje zvuku, a to buď pulzní či šumové, či smíšené, pozměněného lineární funkcí V(f), která zastupuje průchod zvuku ze zdroje vokálním traktem (vokální trakt sice není lineální, ale lze jej za lineární považovat), a radiační charakteristikou R(f), která vyjadřuje změny zvuku při změně prostředí z vokálního traktu do okolního prostředí (obr. 10).

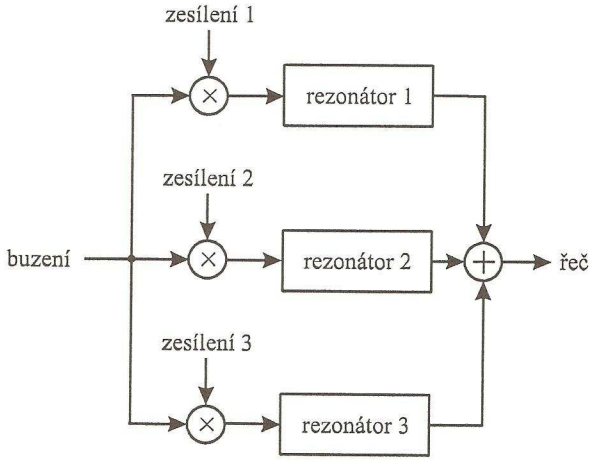


Obr. 10 Model produkce řeči podle teorie zdroje a filtru („z“ odpovídá hodnotám „f“ po z-transformaci) (převzato z Psutka, 1996)

Průchod zvuku vokálním traktem je simulován souborem rezonátorů a antirezonátorů, které zesilují, případně potlačují určité frekvence. Rezonátory mohou být zapojeny buď sériově (kaskádově), jejich funkce se pak násobí, jsou vhodné pro syntézu samohlásek (obr. 11), anebo paralelně, kde se výsledky jejich funkcí sčítají (obr. 12). Paralelní zapojení je vhodnější pro simulaci šumových zvuků, tedy souhlásek. Kvalitní syntetizér by tedy měl obsahovat oba typy s možností přepínání.



Obr. 11 Schéma sériového zapojení rezonátorů (převzato z Psutka, 1996)



Obr. 12 Schéma paralelního zapojení rezonátorů (převzato z Psutka, 1996)

K syntéze se používá pravidel, která udávají jednotlivé hodnoty. Pravidla se dají stanovit buď manuálně na základě souboru poznatků o akustické rovině produkce řeči, či je možné je odhadovat automaticky přímo z velkých korpusů řečových dat. Druhý způsob je však zatím výjimečný. Extrahovaná pravidla (např. délka trvání, hodnoty jednotlivých formantů) se pak nastavují pro každou hlásku syntetizované řeči, přechody mezi hláskami jsou pak doplněny aplikací dalších pravidel (např. tranzienty). Ukazuje se, dostatečná aktualizace parametrů je pro nejrychlejší formantové charakteristiky dostačující kolem 5-10ms.

1.1.2.2.2.1 Výhody a nevýhody

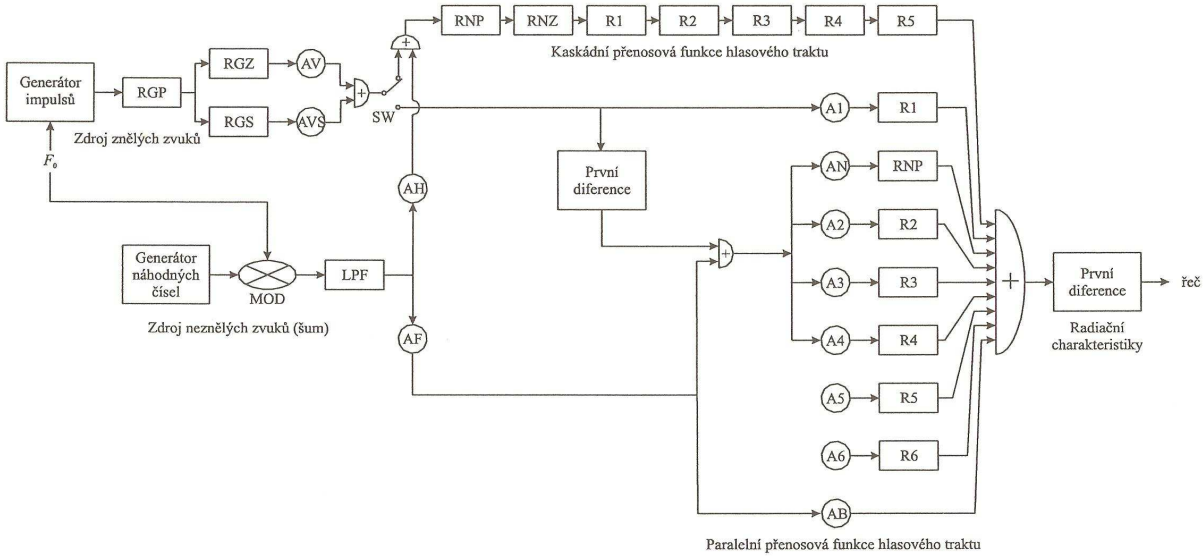
Výhodou formantové syntézy je její výpočetní jednoduchost, počítá se maximálně se 40 parametry. Jednoduše se zde manipuluje s prozodickou charakteristikou řeči a rovněž koartikulace se dá řešit nenáročně, řeč působí plynule. Také tato metoda není závislá na řečníkovi, s barvou hlasu se dá manipulovat. Na druhou stranu, dá větší práci parametry nastavit tak, abychom dosáhli přirozeně znějícího zabarvení a řeč nepůsobila příliš kyberneticky. Navíc hledání pravidel je časově náročné, protože se zatím musí provádět

převážně manuálně. Nicméně na toto manuální hledání se dá nahlížet i jako na přínos k teoretickým znalostem, které o mluvené řeči máme.

1.1.2.2.2.2 Klattovská syntéza

Nejznámější a nejvyužívanější formantovou metodou je metoda Denise Klatta z roku 1980. Je na ní založena i úspěšná metoda DECtalk.

Principem klattovské syntézy je hybridní sériově (kaskádově) paralelní zapojení rezonátorů, resp. jejich počítačové implementace (obr. 13).



Obr. 13 Schéma klattovského syntetizátoru (převzato z Psutka, 1996)

Pro popis zvuku zavádí 40 parametrů - 12 konstantních, charakterizujících typ zvuku (hlasu) (šířka pásma hlasivkového rezonátoru, šířka pásma nosního antiformantu, šířka pásma nosního rezonátoru, frekvence nosního rezonátoru, šířka pásma hlasivkového antirezonátoru, frekvence hlasivkového rezonátoru,…) a celkové nastavení systému (počet formantů v sériové větvi, řízení celkového zesílení, počet vzorků na segment, vzorkovací frekvence) a 28 parametrů proměnlivých v čase, které specifikují změny vlny v průběhu promluvy (amplituda znělých zvuků, amplituda frikativních zvuků, amplituda aspirativních zvuků, základní frekvence, frekvence 1.-6. formantu, amplituda 1.-6. formantu, šířka pásma 1.-6. formantu) Přepínání mezi paralelním a sériovým zapojením zajišťuje hodnota (0/1) parametru SW.

1.1.2.2.3 HL technologie (Stevens, Bickley, 1991)

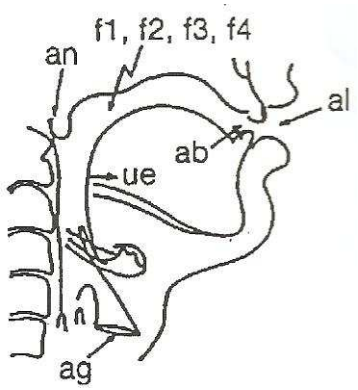
V roce 1991 přicházejí Stevens a Bickley s návrhem vylepšení klattovské syntézy. Poznamenávají, že parametrická syntéza je nejen mocným nástrojem pro syntézu lidské řeči (ovšem v současné chvíli nadlidsky komplikovaným a neintuitivním), ale také dobrým nástrojem pro popis fonetického inventáře – ovšem i v této oblasti by se hodilo, aby byly parametry intuitivnější a jejich souvislost s artikulací zjevnější.

Využili faktu, že plynulá řeč má mnohá omezení a pravidelnosti, které vylučují některé kombinace parametrů a jiné kombinace předurčují – je to zejména koartikulace a aerodynamické a akustické procesy související s tvarem a možnostmi vokálního traktu. Všech 40 parametrů tak není zcela na sobě nezávislých a jisté závislosti je možno dopočítávat automaticky. Protože ona omezení a pravidelnosti vyplývají z artikulační povahy hlásek, celý systém se tak více přiblíží artikulační syntéze, která je intuitivnější.

Zavedli tak 10 tzv. high-level parametrů, k nimž stvořili aparát, který je automaticky mapuje na parametry klattovské, a ty jsou pak syntetizovány.

Byly to tyto parametry (obr. 14):

- f1, f2, f3, f4 – první čtyři formanty reprezentující nastavení tvaru vokálního traktu při zavřeném velu bez konstrikce jazyka či rtů. Tyto parametry jsou přímo mapovány na klatovské parametry prvního až čtvrtého formantu;
- f0 – základní frekvence kmitání hlasivek, v první verzi HL syntézy byl přímo mapován na klatovský parametr základní frekvence;
- ag – otevření glottis, pro znělou řeč se hodnoty pohybují okolo 4mm^2, pro neznělou řeč dosahují 40mm^2.
- al – míra oddálení rtů, hodnota 100mm^2 odpovídá nastavení při nelabiálních hláskách, při úplné konstrikci u bilabiál klesne na nulu;
- ab – nejužší místo konstrikce při konsonantech tvořených čepelí jazyka, hodnota 100mm^2 odpovídá nastavení při hláskách bez konstrikce čepelí jazyka, při úplné konstrikci u alveolár a palatál klesne na nulu
- an – otevření měkkého patra, neboli otevření průchodu do nosní dutiny, nabývá opět hodnot 0-100mm^2, 0mm^2 pro nenazální a nenazalizované hlásky, 100mm^2 pro nazály;
- ue – rozšiřování/zmenšování objemu dutin během závěru u exploziv (více v kapitole Parametr ue – rozšiřování dutin 2.1.1.5)



Obr. 14 Základní parametry systému HLSyn (převzato z Sensimetrics (2004). HLsyn, High-Level Speech Synthesizer User Interface Manual)

V roce 1994 svůj systém vylepšili přidáním dalších tří parametrů:

- ps – subglotální tlak (v cm H_2O);
- dc – poddajnost hlasivek (v procentech); tyto dva parametry zjemňují nastavení průběhu hlasivkové činnosti a tím kvalitu základní frekvence, tedy f0 už není přímo mapována na klattovský parametr základní frekvence;
- ap (area of postrior glottal chink) – udává míru nedovření hlasivek během hlasivkového cyklu, rozsah hodnot je 0-10mm^2. Přínosem tohoto parametru je to, že tak můžeme získat zdroj šumu během znělých částí řeči. V praxi se tento parametr využívá k přidání mírné dyšnosti a tím zpřirozenění syntetizované řeči, protože „čistě počítačová“ řeč působí příliš kyberneticky.

1.1.3 Metoda v naší práci

Pro naši práci budeme využívat HL metody, konkrétně implementaci HLSyn Version 2.2.

1.2 R-ové hlásky (Ladefoged, Maddieson, str. 215-236, 1996)

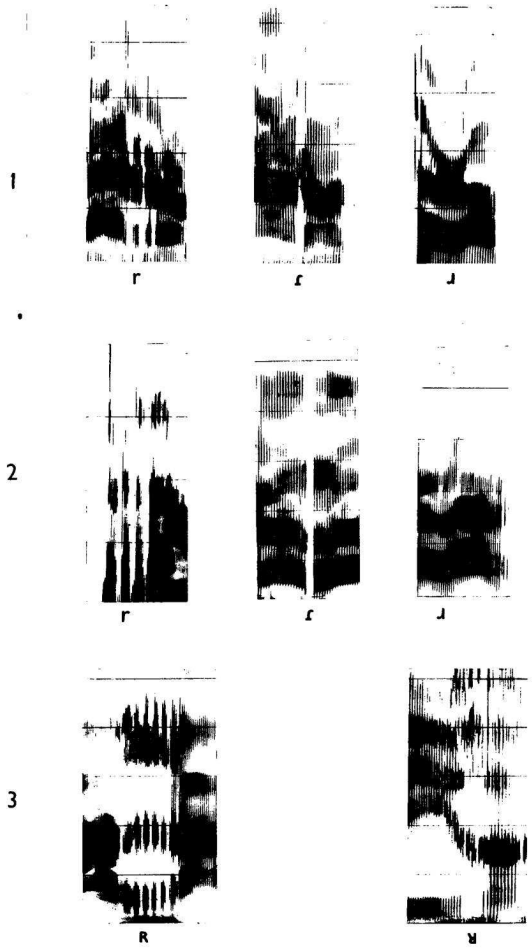
R-ové hlásky, někdy také nazývané anglickým termínem „rhotics“, tvoří hlásky, které výjimečně nesdílejí žádný jeden konkrétní artikulační či akustický rys, nýbrž jsou definovány na základě toho, že se pro ně používá grafému ‚r‘. Rhotics se jim pak říká podle řeckého názvu grafému „ró“ (angl. „rho“).

Ukazuje se, že stanovit jeden fonetický korelát prolínající všemi r-ovými hláskami, je značně složité. Uvažovalo se o sníženém třetím formantu (Ladefoged, 1975, zmíněno v Lindau 1980, bez detailnější reference), ale ani tento směr se neukázal být spolehlivým ukazatelem. Hodnota třetího formantu totiž odráží především místo konstrikce a zaokrouhlení rtů, což je ve většině případů podobné, ovšem ne u všech typů.

Užívání tohoto grafému pro širokou škálu hlásek však zřejmě není náhodné – r-ové hlásky přes svou artikulační odlišnost vykazují podobné fonologické chování – například mají výjimečné postavení ve slabičné struktuře, nezřídka jsou jediným konsonantem, který může stát jako druhý člen konsonantického shluku v prétuře, či jako první člen v kodě. Často mívají r-ové hlásky schopnost stát také v jádře slabiky místo vokálu. Dalším fonologickým vodítkem, proč se pro r-ové hlásky užívá jednoho grafému a je tedy důvod je řadit do jedné fonologické skupiny, je fakt, že r-ové hlásky velmi často alternují mezi sebou, nehledě na místo a způsob tvoření. Například v jazyce farsi má /r/ tři alofony – v iniciální pozici se realizuje jako znělá vibranta, v intervokalické pozici jako švih a ve finální pozici jako neznělá vibranta. (Lindau, 1980)

Vymezením společného rysu r-ových hlásek se zabývá Lindau (1980), která rozborem jednotlivých typů dochází k závěru, že vždy některé spolu sdílejí nějaký rys a ty pak zase s dalšími typy sdílejí jiné své rysy. Klademe si však otázku, proč by toto vymezení mělo stačit ke stanovení fonologické třídy, domníváme se totiž, že touto metodou by pak bylo možno všechny hlásky prohlásit za r-ové hlásky, protože každá hláska sdílí nějaký rys s hláskami z jiné třídy (například znělost/neznělost)

Nejpočetnějšími zástupci r-ových hlásek jsou vibranty tvořené špičkou či čepelí jazyka. Dalšími skupinami r-ových hlásek jsou vibranty s jiným místem tvoření (retroflexivní a uvulární), švihy, frikativní r-ové hlásky a aproximantní r-ové hlásky.



Obr. 15 Spektrogramy jednotlivých r-ových hlásek, řádek 1: alveolární vibranta (španělština), švih (španělština), aleveolární aproximanta (angličtina); řádek 2: alveolární vibranta (švédština), švih (jazyk degema), aleveolární aproximanta (jazyk degema); řádek 3: uvulární vibranta (švédština), uvulární aproximanta (švédština) (převzato z Lindau, 1980)

1.2.1 Frikativní a aproximantní r-ové hlásky

Tyto r-ové hlásky nespočívají v kontaktu artikulátorů, ale v jejich dostatečném přiblížení. Zástupcem je alveolární /r/ v americké angličtině, nebo francouzské uvulární /r/.

České /ř/ spadá částečně i do této kategorie, je na něj nejčastěji pohlíženo jako na frikativní vibrantu, kde frikce a vibrace probíhají zároveň. Na jeho podstatu však různí autoři pohlížejí odlišně (více níže).

1.2.2 Švihy

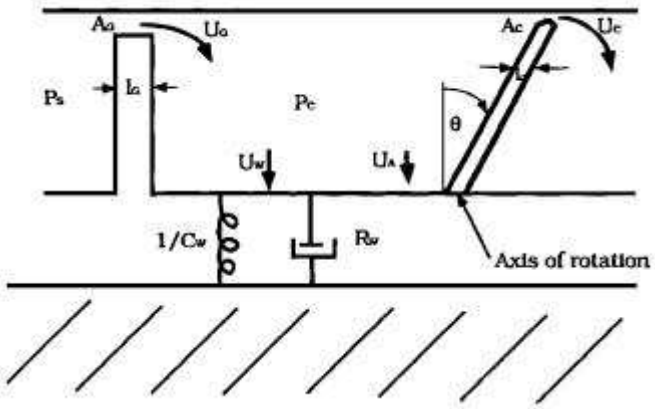
V češtině se užívá termínu „švihy“ pro skupiny označované anglicky tradičně „taps“ a „flaps“. Oba dva typy spočívají ve velmi krátkém kontaktu aktivního artikulačního orgánu s pasívním protějškem. Ladefoged a Maddieson (1996) předkládají následující distinkci: „flap“ jsou hlásky tvořené krátkým kontaktem mezi artikulátory, který vzniká jako důsledek pohybu směrem k místu dotyku, zatímco „taps“ vznikají při pohybu aktivního artikulátoru směrem přímo k patru.

Flaps jsou většinou tvořeny stažením jazyka za alveoly, při jehož návratu dojde ke kontaktu s alveolami, taps nejčastěji vznikají jako přímé přiblížení k dentální či alveolární oblasti.

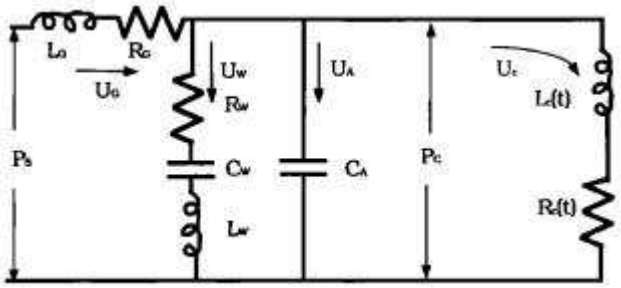
Závěrová fáze švihů je průměrně dlouhá 20ms.

1.2.3 Vibranty

Vibranty jsou charakterizovány jako hlásky, jež vznikají přibližováním a oddalováním dvou artikulačních orgánů způsobeným aerodynamickými podmínkami. McGowan (1992) vysvětluje princip tímto příměrem: „špička jazyka se dá přirovnat k padacím dveřím; ústní dutina k prostoru [pod dveřmi], kde je větší tlak než atmosférický; rozdíl tlaků způsobí, že se dveře otevřou (…), vzduch unikne z prostoru pryč a tlaky se vyrovnají, takže se dveře opět zavřou; prostor je opět naplněn vzduchem přicházejícím z plic (…) a proces cyklicky pokračuje.“ Tento princip, nazývaný Bernoulliho princip, je velmi podobný principu vibrací hlasivek. Podstatné na něm je, že vibrace není způsobena žádnými svaly artikulačních orgánů. McGowan stvořil matematický model tohoto principu, uvádíme nákres a korespondující diagram elektrického obvodu (obr. 16, obr. 17).



Obr. 16 Model principů vibrací, Ps – subglotální tlak, Pc – intraorální tlak, Ug – objemová rychlost v glottis, Uc – objemová rychlost v místě konstrikce špičkou jazyka, Uw – objemová rychlost stěn vokálního traktu, Ua – objemová rychlost způsobená soudržností vzduchu, Ag – místo glottálního průchodu, Ac – místo průchodu konstrikcí, lg – tloušťka glottis, lc – tlouštka špičky jazyka, theta – úhel rotace špičky jazyka, Cw – soudržnost stěn vokálního traktu, Rw – odpor stěn vokálního traktu (převzato z McGowan, 1992)



Obr. 17 Model elektrického obvodu simulujícího princip vibrací, Ps odpovídá celkovému napětí, které je do obvodu přiváděno, C jsou kondenzátory a jejich kapacitance, R jsou rezistory a jejich odpor, L jsou cívky a jejich induktance (převzato z McGowan, 1992)

Dochází zde tak k podobnému konfliktu jako u definice znělosti mezi akustickým vymezením, které se zakládá na vícečetnosti vibrací, a artikulačním vymezením, které vyžaduje dostatečné přiblížení orgánů tak, aby mohla vibrace v důsledku Bernoulliho efektu vzniknout. Ačkoliv by se mohlo zdát, že jde o to samé, není tomu tak – lze to ilustrovat na příkladu českého /r/, respektive jeho jednokmitné varianty (která, jak se zdá, je v běžné spontánní řeči zastoupena nejvíce), které se řadí mezi vibranty, ačkoliv

k vícečetnosti kmitů (akustické hledisko) nedochází. Podstatné je, že na jejím vzniku se podílí ten samý princip (dostatečné přiblížení orgánů a vyrovnávání tlaků), který je zodpovědný za vibrace (artikulační hledisko).

Z hlediska náročnosti produkce je nejvýhodnější, když je rozkmitávána relativně malá masa artikulačního orgánu, nejčastější jsou tak vibranty tvořené špičkou jazyka kmitající v dentální či alveolární oblasti, nebo čípkem kmitajícím proti kořeni jazyka. Jiná místa tvoření jsou poměrně vzácná, jako například české /ř/, u nějž kmitá čepel jazyka.

Vibranta se tak skládá z 1 a více cyklů vibrací. Takovýto cyklus zahrnuje fázi, kdy jsou orgány přiblíženy, a fázi oddálení. Každá přiblížení jsou oddělena jednou mezifází, která má pseudovokalické vlastnosti. Jedna vokalická mezifáze trvá okolo 25ms, stejně tak jedna fáze doteku. Vibrace tak mají frekvenci přibližně 20Hz. Před prvním dotekem se nachází podobná pseudovokalická fáze jako mezi dvěma doteky, stejně po posledním doteku. Tyto otevírací a zavírací fáze mají dvojnásobné trvání, tedy okolo 50ms. Pokud má vibranta více kmitů, bývá první krmit delší než následující. (Ladefoged, Maddieson, 1996)

1.2.3.1 Česká alveolární vibranta /r/

1.2.3.1.1 Místo tvoření

Alveolární místo tvoření patří mezi nejčastější, pro popis českého /r/ tedy platí výše uvedené u vibrant.

1.2.3.1.2 Průběh – fáze, počet kmitů, trvání

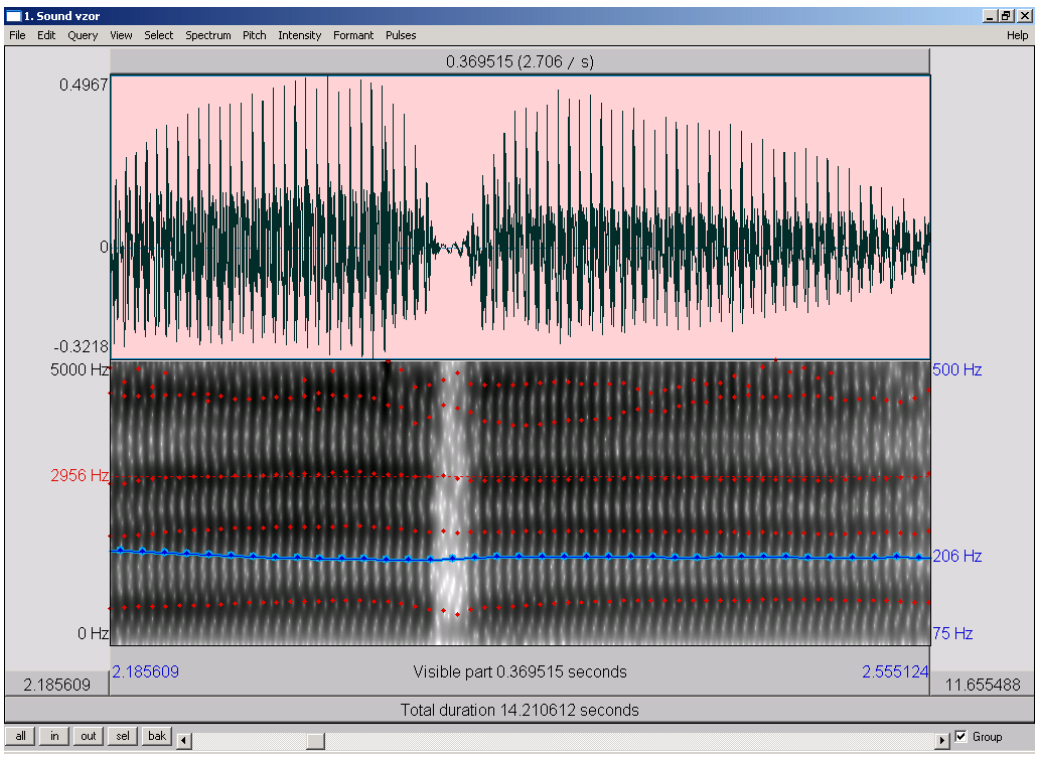
Autoři se různí v názoru na počet kmitů českého /r/, Chlumský (1911) uvádí 1-7 vibrací, Maláč a Borovičková (1967) 1-4, Palková (1994) 1-3 vibrace. Podle nejnovějších studií (Machač, 2009) je české /r/ většinou jednokmitné. Nabízí se otázka, zda se tedy nemůže jednat o švih. Podle Recasense (1991) mají ovšem vibranty a švihy zcela odlišné akustické vlastnosti, stejně tak odlišné chování ve vztahu k okolním hláskám. Rozhodně nelze říci, že vibranta je série několika švihů. Je tedy nutné na české /r/ pohlížet skutečně jako na vibrantu s jedním cyklem.

Průměrné trvání uvádějí Maláč a Borovičková (1967) 138ms, frekvenci kmitání 20Hz . Trvání hlásky /r/ se liší podle toho, zda se nachází v pozici iniciální, mediální, či finální – 129ms, 85ms, 198ms. Vzhledem ke krátkému trvání je v mediální pozici většinou jednokmitné.

Pro úplnost opakujeme hodnoty, které udávají Ladefoged a Maddieson: (1996) trvání vokalické mezifáze okolo 25ms, stejně tak fáze doteku, tedy frekvence vibrací přibližně 20Hz, trvání pseudovokalická presekvence a postsekvence okolo 50ms, trvání prvního kmitu delší než následujících.

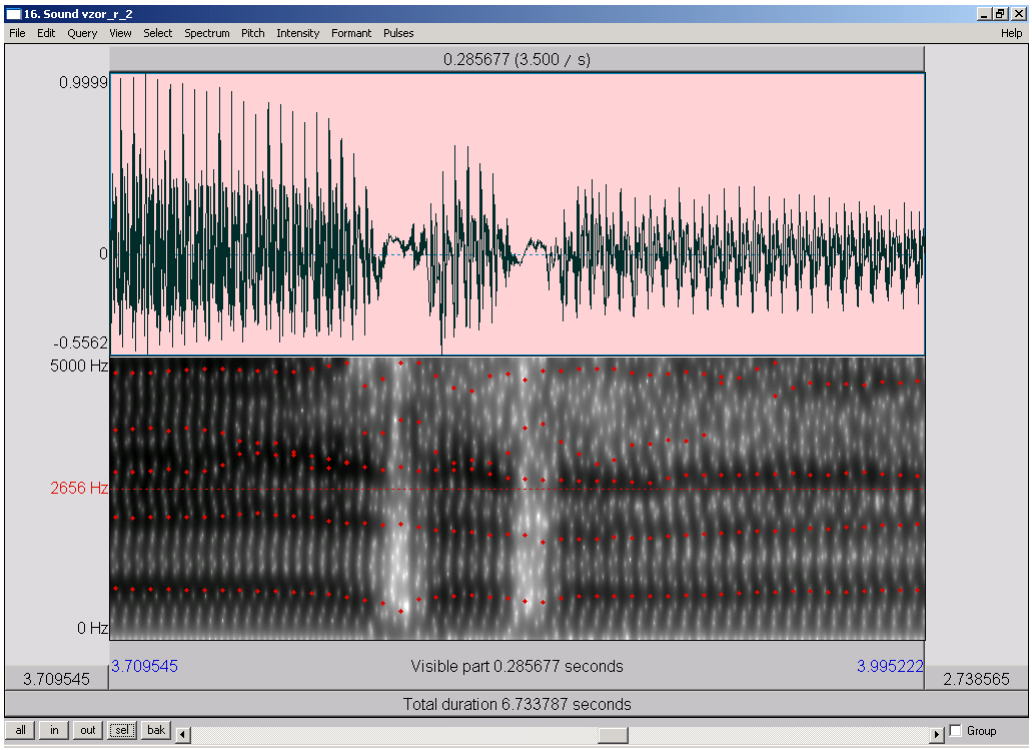
Pozornost průběhu českého vibrantního /r/ věnoval Machač (2009) v rámci práce o segmentaci hlasového materiálu. Jak již uvádí Ladefoged Maddieson (1996), vibranta se skládá z fází přiblížení a oddálení a fáze oddálení artikulátorů (včetně fáze před prvním přiblížení a fáze po posledním přiblížení) mají pseudovokalickou povahu. Pseudovokalické fáze mohou mít buď kvalitu schwa, nebo přejímat formanty okolních vokálů. Pokud tedy /r/ stojí v intervokalické pozici, je náročné jej spolehlivě oddělit od okolních vokálů.

Pro ilustraci výše uvedeného přidáváme spektrogram jednokmitného a vícekmitného /r/ (obr. 18 a 19).



Obr. 18 Jednokmitné /r/ v intervokalickém okolí (/ere/)

Na spektrogramu je patrné splývání pseudovokalické presekvence a postsekvence s okolními vokály.



Obr. 19 Dvojkmitné /r/ v intervokalickém okolí (/ere/)

Na spektrogramu je vidět mezikmitná pseudovokalická fáze.

1.2.3.1.3 Formantová struktura

Maláč a Borovičková (1967) uvádějí průměrnou hodnotu prvního formantu 600Hz, druhého formantu 1600Hz a třetího formantu 2500Hz.

My vycházíme z hodnot pro alveolární explozivy.

1.2.3.2 Česká (post)-alveolární vibrantní frikativa /ř/

1.2.3.2.1 Místo tvoření

Ladefoged a Maddieson uvádí české /ř/ jako jedinou skutečně se objevující se vibrantu artikulovanou nikoliv špičkou, ale čepelí jazyka.

1.2.3.2.2 Průběh – fáze, počet kmitů, trvání

Chlumský ve své studii z roku 1911 uvádí několik poznatků ohledně akustických i artikulačních vlastností /ř/, z nichž některé lze využít v naší práci. Na hlásku /ř/ lze pohlížet dvěma způsoby – jako na hlásku jednoduchou, kde vibrace i šum probíhají současně, nebo jako na hlásku složenou, kde vibrace a šum přicházejí postupně. Chlumský zastává první názor, tedy že šum je přítomen po celou dobu trvání hlásky, na rozdíl od vibrací, které mohou být přítomny celou dobu, nebo jen v začátku. K této problematice se později vyjadřují další autoři s různými názory – Trávníček (1932) považuje /ř/ za hlásku jednoduchou, stejně tak Kučera (1961). Isačenko (1965) ji naopak považuje za hlásku složenou, Palková (1994) nazývá /ř/ šumovou vibrantou, tedy ji považuje za hlásku jednoduchou.

Dále Chlumský uvádí, že /ř/ je až o třetinu delší ve svém trvání než /r/. Pokud se v druhé části hlásky nevyskytují vibrace, je pouze šumová část obvykle kratší než část vibrantně-šumová. Uvádí, že /ř/ má 2-11 kmitů, Palková (1994) uvádí 2-6 kmitů. Na základě dosud nepublikovaných pozorování existuje dnes předpoklad, že i /ř/ je, podobně jako /r/, v češtině převážně jednokmitné, což potvrzuje i naše nahrávka.

Maláč a Borovičková (1967) uvádějí ohledně trvání následující hodnoty: šumová složka nastupuje v průměru 36ms po vibrační a je dlouhá průměrně 195ms. Celkové trvání hlásky /ř/ je 167ms, 158ms a 260ms pro iniciální, mediální a finální pozici, znělé a neznělé nerozlišují.

1.2.3.2.3 Formantová struktura

Podle Maláče a Borovičkové (1967) je první formant periodické složky průměrně 600Hz, druhý 1600Hz a třetí 2000Hz. Šumová složka má spektrální těžiště v rozmezí 2500-8000Hz. My opět vycházíme z hodnot pro alveolární explozivy.

2 Metoda práce

2.1 Syntéza

Pro parametrickou syntézu jsme vycházeli z několika hledisek (hledisko teoretické a hodnoty získané z reálných dat), které jsme různě zkombinovali, kvalitu výsledných syntetizovaných hlásek jsme pak ověřili percepčním testem, který se zaměřoval na vnímání přirozenosti hlásek laickými posluchači. Hlásky jsme syntetizovali v intervokalickém prostředí, mezi vokály byl zvolen vokál /e/, a to z důvodu nelabialnálnosti (vypadlo /u/, /o/), z důvodu široké slovní zásoby obsahující sekvenci /ere/ i /eře/ (vypadlo /a/, které je v sekvenci /aře/ vzácné), a z důvodu středních hodnot f1 (vypadlo /i/, /í/).

Pro neznělé /ř/ pochopitelně intervokalické prostředí není možné, bylo proto syntetizováno v okolí eř0.

2.1.1 Jednotlivé parametry cílových hlásek pro syntézu

2.1.1.1 f0 – základní frekvence (Hz)

Základní frekvenci jsme nastavili na 120Hz, což odpovídá průměrnému mužskému hlasu (Palková, 1994, str. 57 uvádí rozsah běžných mužských hlasů 100-150Hz).

V HLSynu jsou hodnoty uváděny v desetinásobku, zřejmě kvůli zjednodušení zanesení do grafického znázornění společně s formanty.

2.1.1.2 Formanty f1-f2-f3 (Hz)

Hodnota formantů – tranzientů je závislá na okolních hláskách a na místě tvoření cílové hlásky. Pro alveolární místo tvoření a vokál /e/ se hodnoty druhého formantu pohybují okolo 1800Hz, třetího okolo 2700Hz.

Tranzienty prvního formantu se odvíjejí od způsobu i místa tvoření. Při aleveolární okluzi se hodnota pohybuje okolo 400Hz, pro frikativy je hodnota vždy vyšší než korespondující okluzivní. Pro alveolární frikativy je to okolo 420Hz, což bude důležité při syntéze /ř/, které je považováno za frikativní vibrantu.

Hodnoty druhého formantu jsou závislé na místě tvoření, čím nižší, tím blíže rtům (pro bilabiální je hodnota přibližně 1550Hz, pro alveolární 1900Hz, pro velární 2300Hz). Budeme tedy experimentovat s touto hodnotou, provedeme syntézu pro 1800Hz, 1900Hz, 2000Hz a 2200Hz. Naše hypotéza je, že hodnota 1800Hz platná pro anglické alveoláry, je pro české alveoláry „příliš vepředu“, takže přirozeněji budou znít vyšší varianty.

Maláč a Borovičková (1967) uvádějí průměrnou hodnotu prvního formantu 600Hz, druhého formantu 1600Hz a třetího formantu 2500Hz, vytvoříme i varianty s těmito hodnotami.

Tranzienty u alveolár většinou začínají 50ms před cílovou hláskou a doznívají 50ms po ní, toho se budeme držet stabilně ve všech variantách. Nabízí se otázka, zda nějak upravovat hodnoty formantů během jednotlivých vibrací hlásky /r/. Z pozorování reálného signálu nám vychází, že jsou přechody natolik rychlé, že pokud ke změnám dochází, jsou zanedbatelné.

Pro hlásku /ř/ uvádějí Maláč a Borovičková hodnoty formantů 600Hz, 1600Hz a 2000Hz.

2.1.1.3 Parametr ag – míra otevření glottis (0-40mm^2)

Pro vokály a znělé explozivy je hodnota stabilně 4mm^2. Pro znělé frikativy je 8mm^2. Pro neznělé se hodnoty pohybují od 20mm^2 (explozivy) po 30mm^2 (frikativy).

U přechodu na explozivu se hodnota začíná měnit 15ms před začátkem závěru, zpět zavírat se začíná pak při explozi a je dokončena 20 ms po ní.

U přechodu na frikativu, jsou to časy 50ms před začátkem frikativy a 50ms po jejím skončení. Maximální hodnotu (8/30mm^2) má frikativa ve své polovině, její hranice jsou dopočítané automaticky.

2.1.1.4 Parametr ab – konstrikce špičky/čepele jazyka (0-200mm^2)

Pokud se špička/čepel jazyka při artikulaci nezapojuje, je hodnota 100mm^2, pokud tvoří úplnou konstrikci, klesne až na 0mm^2. Klesání a stoupání trvá u exploziv cca 25ms před, resp. po cílové hlásce. U frikativ je přechod plynulejší a hodnota neklesá až na nulu, nýbrž na hodnoty okolo 7mm^2. Budeme experimentovat především s dobou

zavírání a otevírání, protože ta je pravděpodobně pro vibranty způsobované Bernoulliho efektem charakteristická

2.1.1.5 Parametr ue - rozšiřování dutin (-200-200cm^3/s)

Během závěru znělých exploziv je k činnosti hlasivek (vytvoření znělosti) zapotřebí výdechového proudu vzduchu z plic, ovšem kvůli závěru nemá tento vzduch možnost unikat do okolí a hromadí se v dutinách za překážkou. Jak ukázal J. R. Westbury (1983) pomocí cineflurografie hrtanu, měkkého patra a pozice jazyka, během tohoto závěru dochází logicky ke zvyšování intraorálního tlaku, který má za následek kompresi artikulačních orgánů a tím rozšíření dutin. Změna velikosti dutin se pak odrazí i na akustické kvalitě závěrové fáze exploziv.

V HLSynu je tento jev reprezentován parametrem ue, udaným v cm^3/s.

Protože při syntéze /r/ vycházíme z parametrů pro /d/, je v syntéze tento parametr zahrnut. Nastává otázka, zda bude zapotřebí tento parametr aplikovat i pro rychlé závěry při syntéze vibranty, neboli jestli k rozšiřování dutin dochází i u vibrant. Pokud je nám známo, touto otázkou se teoreticky zatím nikdo nezabýval. Budeme tak ve svých úvahách vycházet z artikulačního principu vibrant.

Jak již bylo řečeno, základní princip vzniku vibrace je založen na přiblížení aktivního artikulátoru k pasivnímu do té míry, že proudícímu vzduchu nevytváří překážku ve smyslu konstrikce u exploziv, ale je tímto vzduchem rozkmitáván. Kmity tak nejsou aktivním svalovým pohybem artikulátoru. Domníváme se tedy, že pro vznik takového přetlaku, aby docházelo k rozšiřování dutin by bylo potřeba, aby byla překážka držena na místě záměrně. Z fyzikálního hlediska je pravděpodobně jednodušší oddálit relativně malou masu artikulátoru, který stojí proudu vzduchu v cestě, než stlačit mnohem pevnější části, které rozšiřování dutin podléhají.

Druhým argumentem je rychlost rozšiřování dutin. Podle autorů HLSynu trvá přibližně 15ms než hodnota parametru ue dosáhne z 0 na 150 pro explozivy, kde je závěrová doba okolo 80ms. V případě 10-25ms závěru u vibranty by tak rozšiřování muselo probíhat výrazně rychleji, což vylučují vlastnosti orgánů, nebo by se hodnota parametru zvedla pouze na poměrnou část hodnoty, což je možné zanedbat, protože by to pravděpodobně nemělo percepčně žádný vliv.

Pro syntézu vibrant tedy parametr ue vynecháváme.

2.1.1.6 Parametr ap – nedovření hlasivek

Parametr ap (míra nedovření hlasivek) jsme stabilně nastavili na hodnotu 5mm^2, aby zvuk získal mírnou dyšnost a tím větší přirozenost.

2.1.1.7 Trvání

Pro /r/ budeme vycházet z průměrného trvání explozivy (80ms), což koresponduje s pozorováním Maláče a Borovičkové (85ms v mediální pozici) a budeme jej upravovat podle L#M (trvání vokalické mezifáze okolo 25ms, trvání doteku 25ms, trvání pseudovokalické presekvence a postsekvence okolo 50ms).
Je však potřeba se ptát, podle jakých pravidel autoři manuálu a Maláč a Borovičková při stanovování délky hlásky /r/ vycházeli – zda započítávali i pseudovokalické presekvence a postsekvence, případně jakou část z nich. Podle délky se zdá, že nějakou část započítávali určitě, protože na samotný závěr by bylo 85ms příliš dlouhých, a protože pseudovokalické části neřešíme (viz Syntéza /r/ 2.1.3), nebudeme se délkou v tomto případě více zaobírat.
Podle reálných dat máme podezření, že jsou fáze ještě kratší, než autoři udávají, provedeme tedy i syntézu v souladu s tímto pozorováním.

Pro /ř/ budeme vycházet mimo jiné z údajů, které uvádí Maláč a Borovičková - šumová složka nastupuje v průměru 36ms po vibrační a je dlouhá průměrně 195ms.

2.1.2 Syntéza vokálu /e/

Pro syntézu vokálu jsme zvolili následující hodnoty (f1, f2 podle Volína, 2007, f3, f4 základní nastavení HLSynu)

f1 = 620
f2 = 1700
f3 = 2500
f4 = 3500

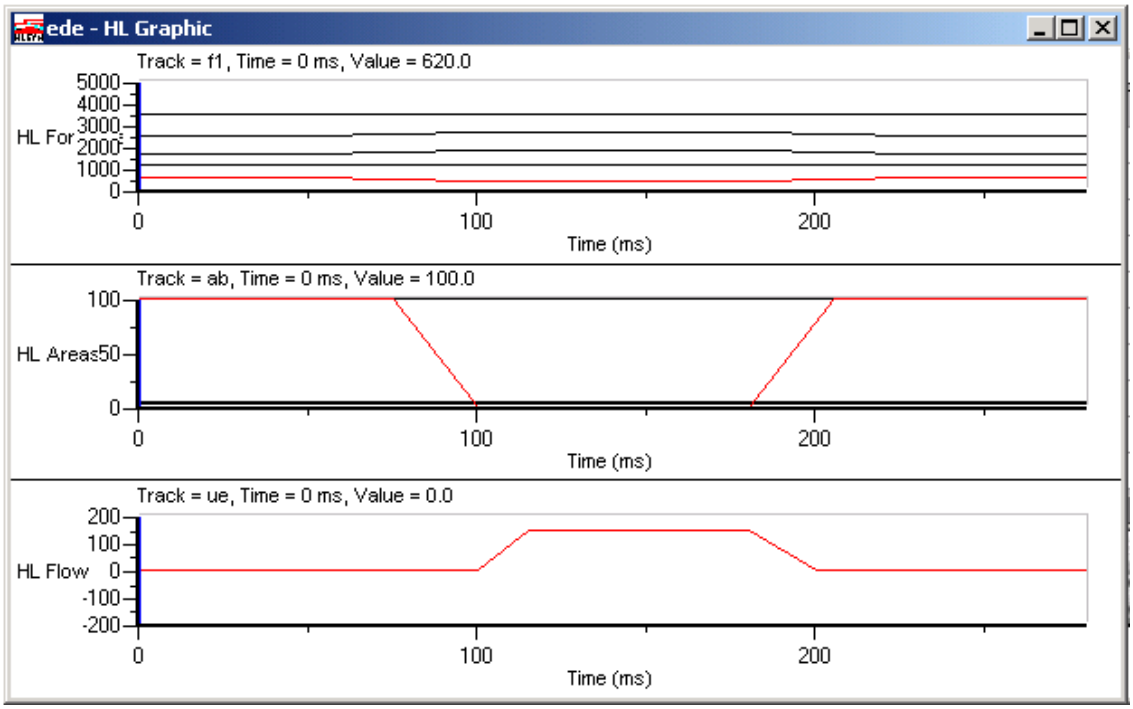
Trvání vokálu je vždy shodně 100ms v mediální pozici, 140ms ve finální pozici.

2.1.3 Syntéza /r/

Pro všechny pokusy platí, že předpokládáme, že pseudovokalické presekvence a postsekvence mají dle Machače (2009) formantovou strukturu velmi blízkou sousedním vokálům, takže s vokály splývají a jejich syntézu neřešíme. (viz 1.2.1.1 Česká alveolární vibranta /r/)

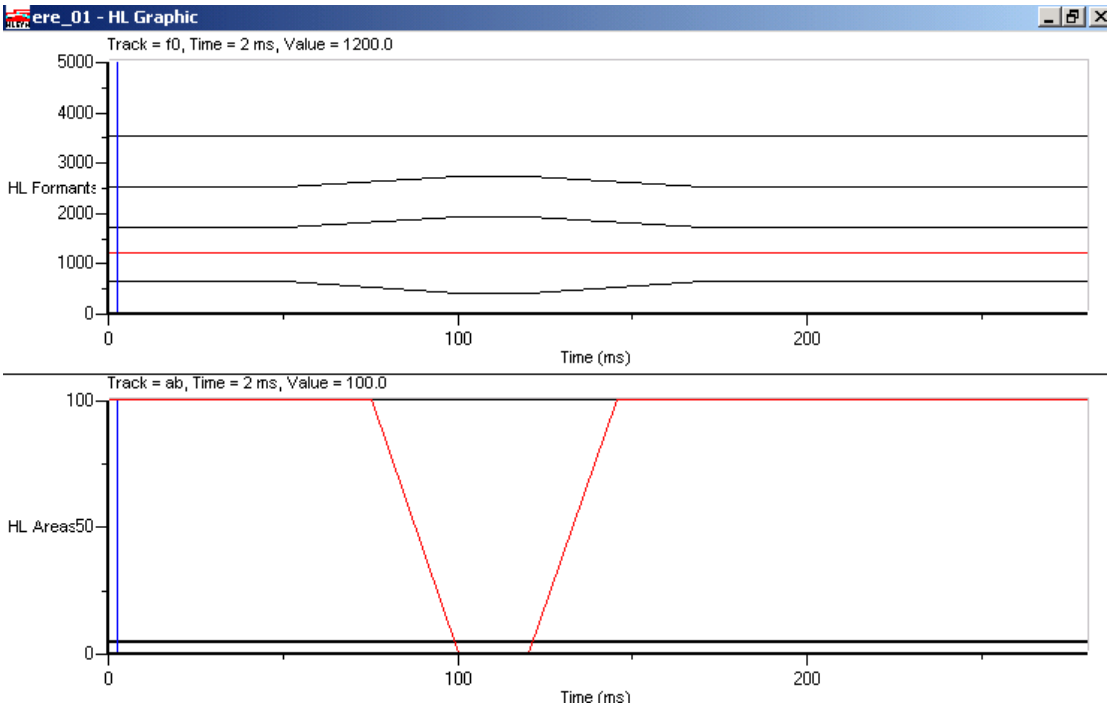
2.1.3.1 Pokus první – zkrácení závěru /d/ (ere_01)

Jako první jsme syntetizovali hlásku /r/ jako velmi krátkou explozivu s odpovídajícím místem artikulace, tedy se stejnými parametry jako pro hlásku /d/. Trvání závěru v sekvenci /ede/ bylo stanoveno na 80ms.



Graf 1 Parametry syntézy sekvence /ede/

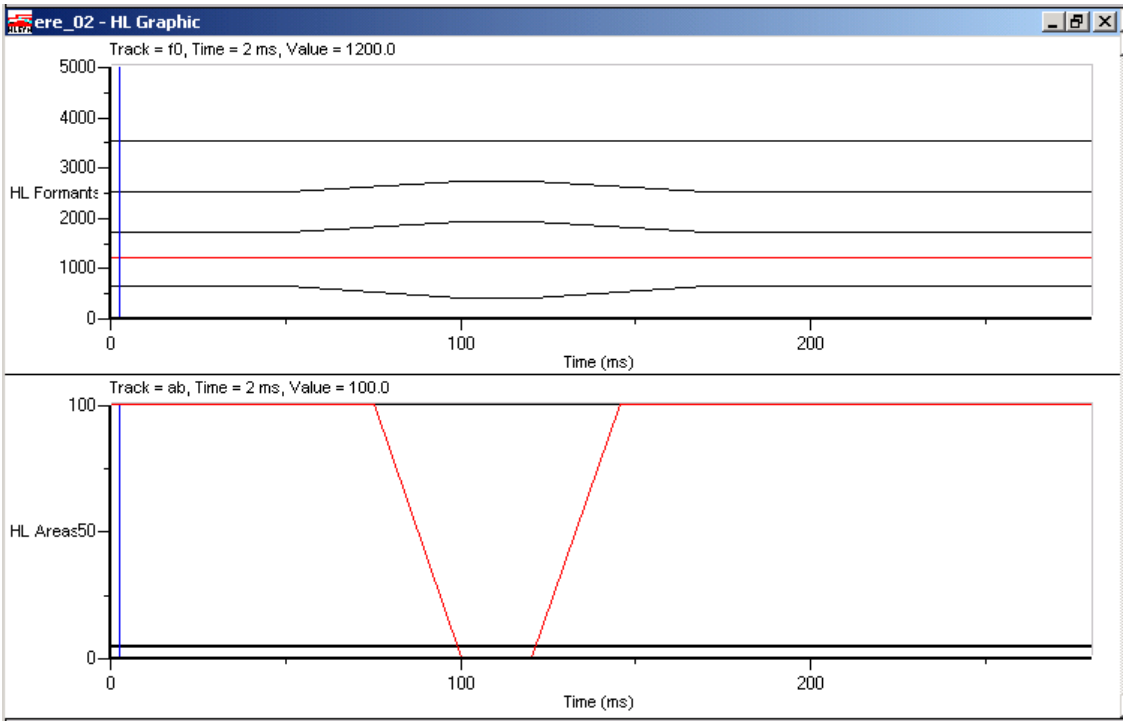
Pro sekvenci /ere/ jsme trvání závěru zkrátili na 25ms (podle Ladefogeda, Maddiesona, 1996). Parametr ue jsme v souladu s naší úvahou (Parametr ue – rozšiřování dutin 2.1.1.5) vypustili. Samozřejmě se zkrácenou dobou závěru jsme upravili i hodnoty přechody formantů.



Graf 2 Parametry syntézy sekvence /ere/, zkrácení závěru (ere_01)

2.1.3.2 Pokus druhý – posun krátkého /d/ dozadu (ere_02)

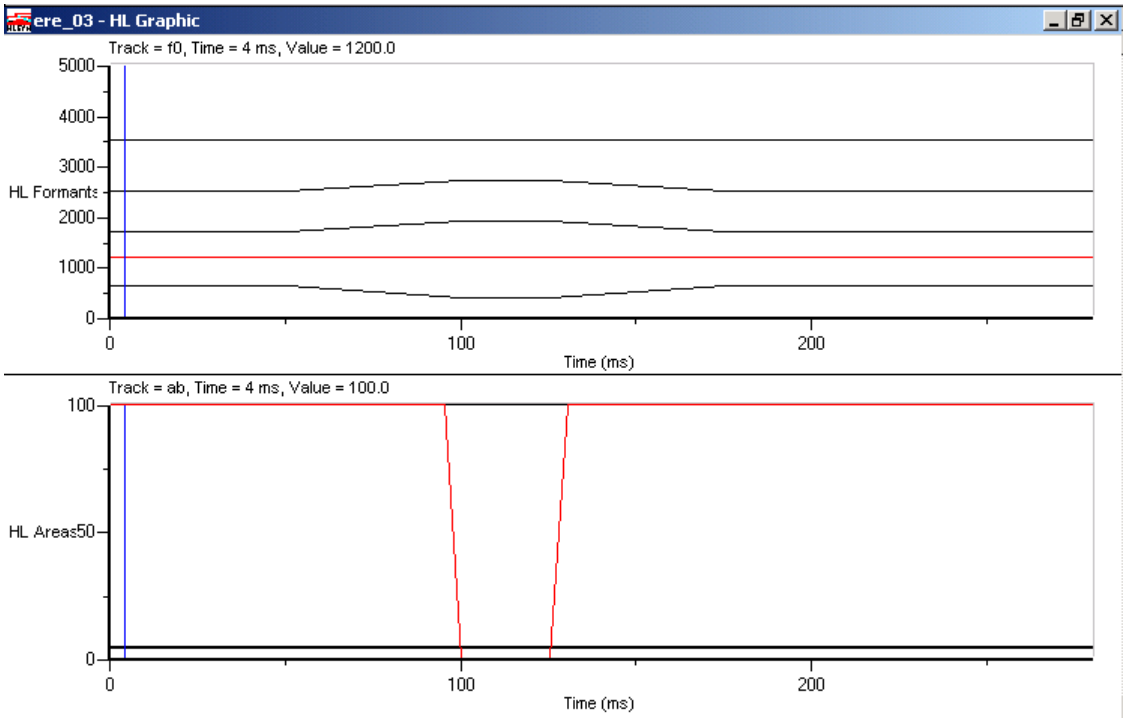
Oproti tabulce tranzientů z manuálu jsme f2 nastavili na 1900 (místo 1800), protože české /d/ je oproti anglickému posunuté více na alveoly.



Graf 3 Parametry syntézy sekvence /ere/, posun dozadu (ere_02)

2.1.3.3 Pokus třetí – změna rychlosti uzavírání a otevírání čepele jazyka (parametr ab) (ere_03)

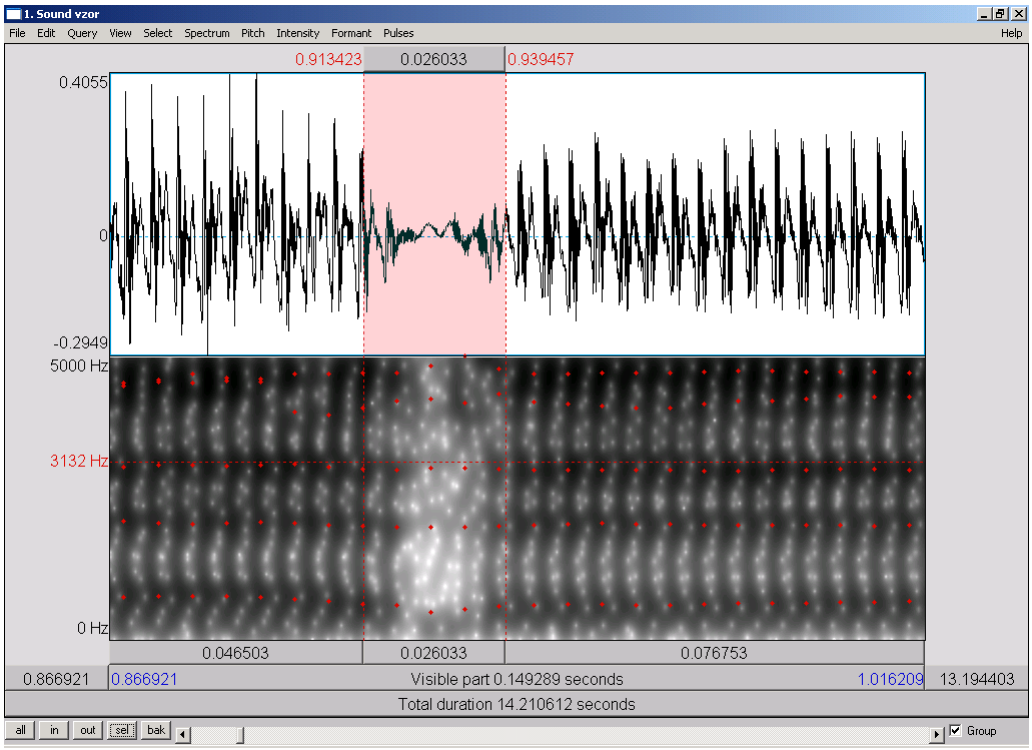
Stále zůstáváme u jednokmitné varianty, budeme si více všímat parametru ab. Protože jsou kmity u /r/ rychlejší a nejsou způsobovány aktivní činností artikulátorů, nýbrž aerodynamickými podmínkami (Bernoulliho efekt), mělo by být rychlejší i zavírání a otevírání. Čas otevírání jsme tak zkrátili z 25ms na 7ms. Protože se předpokládá, že nástup a ústup akustické energie má tendenci být symetrický (Machač, 2009, str. 179), čas zavírání jsme nastavili stejně, tedy na 7ms.



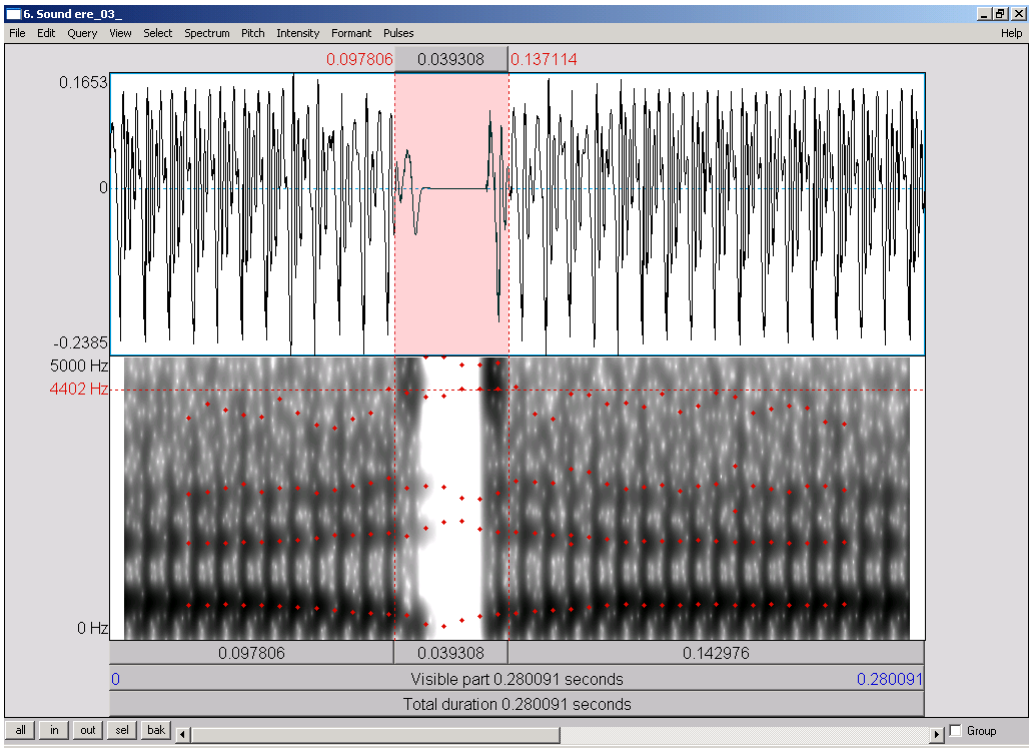
Graf 4 Parametry syntézy sekvence /ere/, zkrácená doba zavírání a otevírání čepele jazyka (ere_03)

2.1.3.4 Pokus čtvrtý – větší zkrácení doby závěru (ere_04)

Srovnáním spektrogramu reálného jednokmitného /r/ v intervokalické pozici (vlastní nahrávka) a spektrogramu syntetizovaného /r/ jsme zjistili (obr. 20 a 21) , že času 25ms odpovídá v reálu celá závěrová fáze včetně doby zavírání a zavírání, nikoliv pouze doba závěru. Ta má v reálném spektrogramu pouze 10ms.

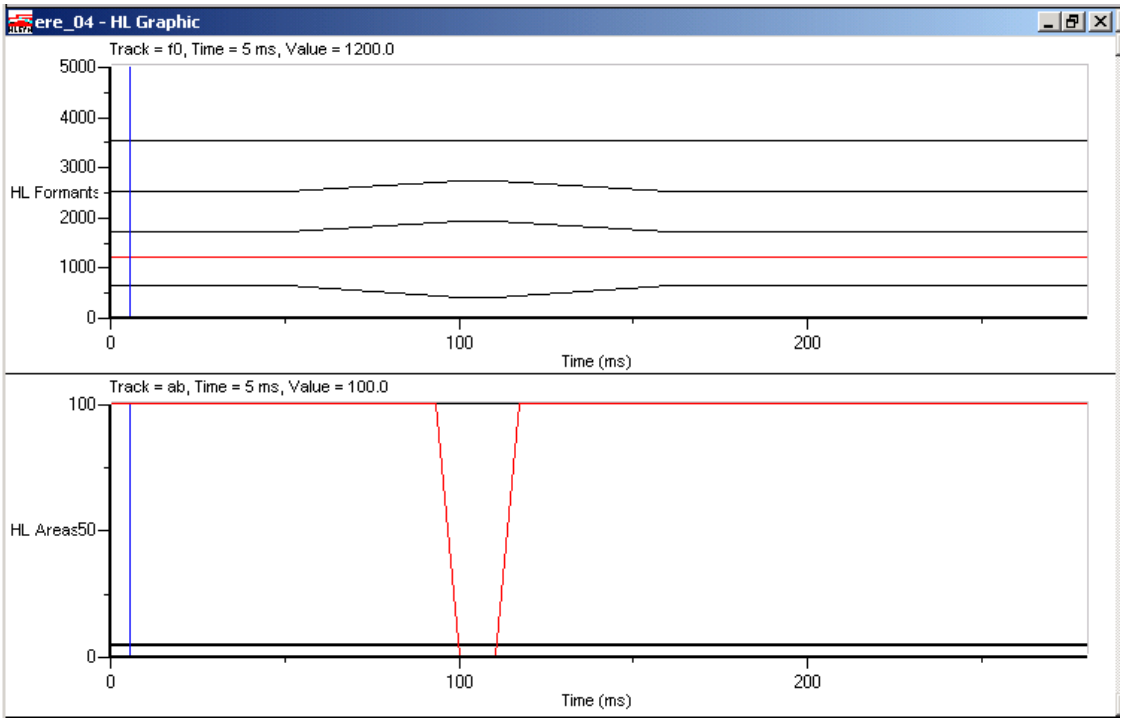


Obr. 20 Trvání závěrové fáze /r/ v reálném zvuku, včetně fáze zavírání a otevírání



Obr. 21 Trvání závěrové fáze /r/ v syntetizovaném zvuku, včetně fáze zavírání a otevírání

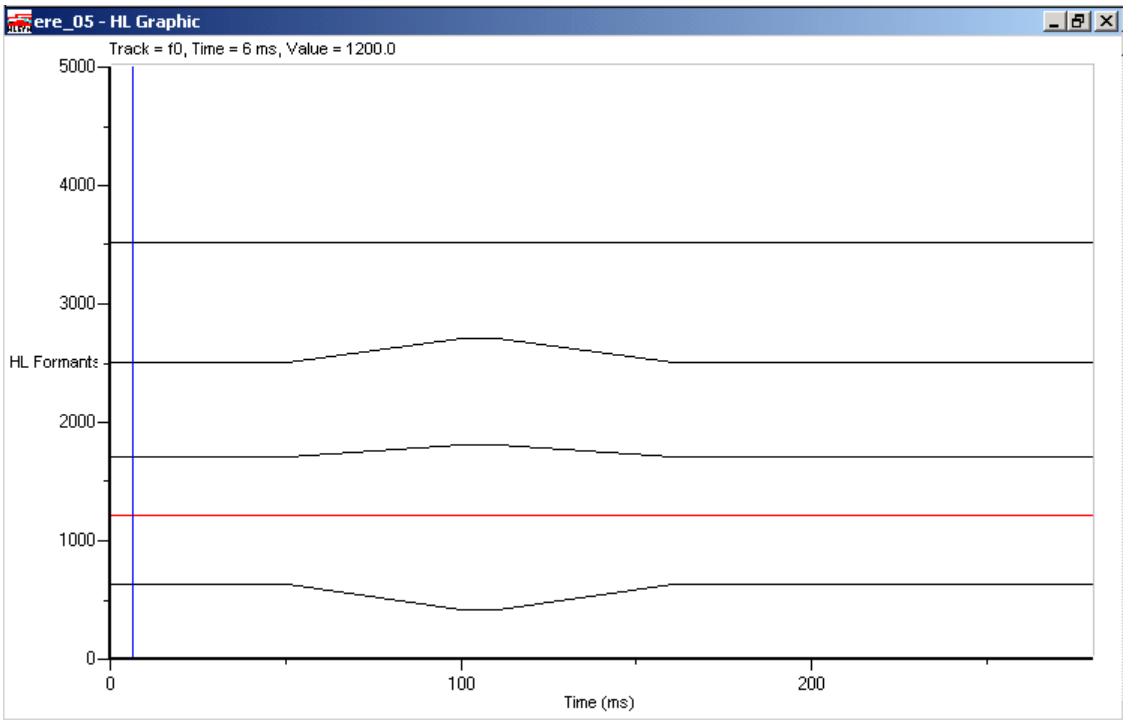
Další varianta tedy nastavuje parametr ab tak, že závěr trvá 10ms, a začíná i odeznívá 7ms před, respektive po.



Graf 5 Parametry syntézy sekvence /ere/, zkrácená doba závěru i zavírání a otevírání čepele jazyka (ere_04)

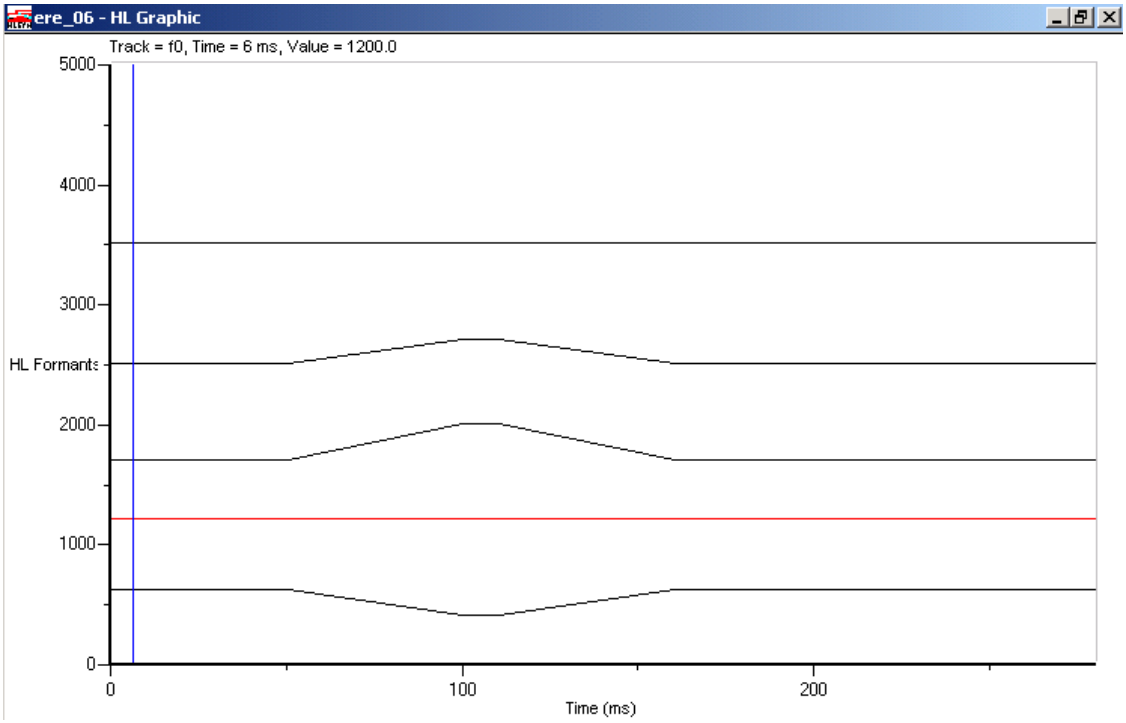
2.1.3.5 Pokus pátý – f2: 1800Hz (ere_05)

Abychom se dozvěděli více o místu tvoření /r/, syntetizovali jsme jednokmitnou variantu i pro další hodnoty f2. První je snížená hodnota na 1800Hz, tedy mírný posun dopředu. (Tato hodnota byla již u úplně prvního pokusu, ale ten byl ještě značně nedokonalý v jiných ohledech.)



Graf 6 Parametry syntézy sekvence /ere/, posun dopředu (ere_05)

2.1.3.6 Pokus šestý – f2: 2000Hz (ere_06)



Graf 7 Parametry syntézy sekvence /ere/, posun dozadu (ere_06)

2.1.3.7 Pokus sedmý – výrazný posun dozadu (ere_07)

Posledním posunem je hodnota druhého formantu na 2200Hz.

2.1.3.8 Pokus osmý - desátý – formanty podle Borovičkové a Maláče u ostatních variant (ere_09 – ere_10)

Protože se hodnoty, které uvádějí Borovičková a Maláč (1967) s našimi hodnotami rozcházejí, vytvořili jsme i verze, které respektují jejich naměřené hodnoty formantů, čili 600Hz, 1600Hz a 2500Hz.

Abychom ověřili, zda jsou lépe vyhovující hodnoty f1-f3, ke kterým jsme dospěli přechodem od hlásky /d/, či hodnoty, které uvádějí Maláč a Borovičková, zahnuli jsme do percepčního testu všechny varianty zdvojeně. Odpovídající páry tak jsou:

ere_1 - ere_8 (krátké /d/)

(ere_2 – krátké /d/ posunuté vzad hodnotou formantu f2 je bezpředmětné)

ere_3 - ere_9 (rychlejší zavírání/otevírání)

ere_4 - ere_10 (kratší závěr)

Pokud je jedna varianta lepší, očekáváme stabilně lepší výsledky u všech párů variant.

2.1.3.9 Pokus jedenáctý – třináctý – zjištění zásadního rozdílu mezi sadami formantů (ere_11 – ere_13)_

Abychom odhalili zásadní rozdíl mezi dvěma sadami formantů, zkusili jsme utvořit tři varianty, vycházeli jsme z hodnot ere_04 a vždy jsme změnili jeden z formantů na hodnotu Maláče a Borovičkové.

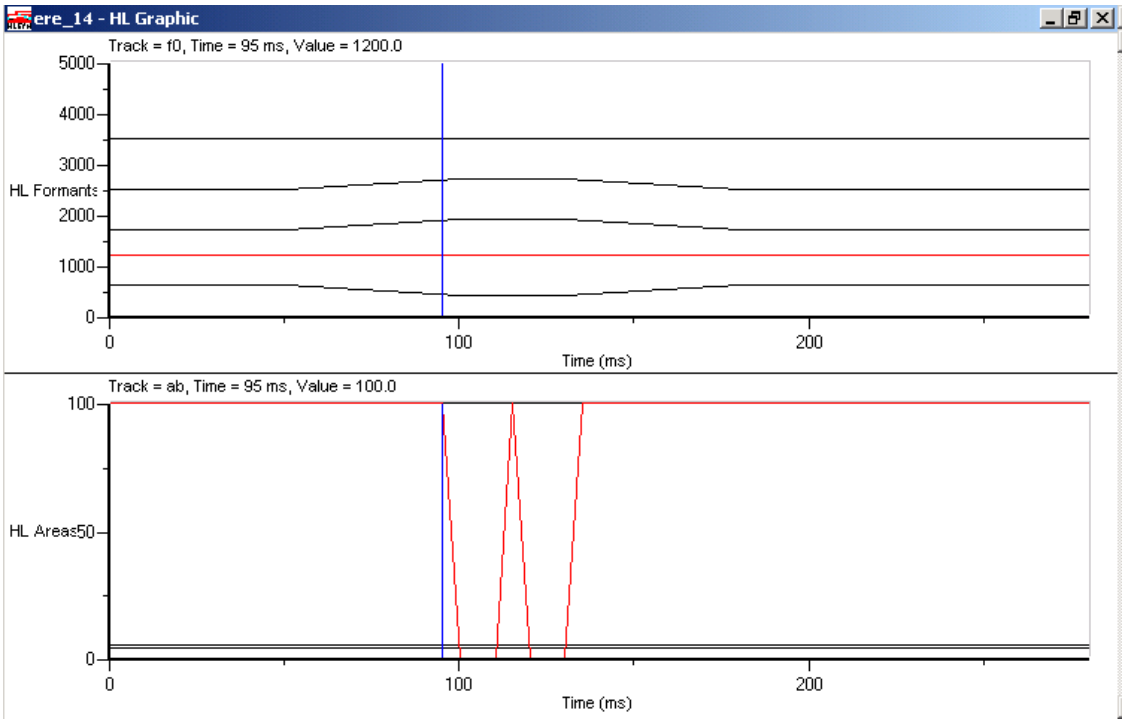
Jedenáctá varianta se tak liší v hodnotě f1 – místo 400 má 600.

Dvanáctá varianta se liší v hodnotě f2 (korelátu místa artikulace) – místo 1800 má 1600 (posun dopředu)

Třináctá varianta se liší v hodnotě f3 – místo 2700 má 2500.

2.1.3.10 Pokus čtrnáctý - dvojkmitná varianta (ere_14)

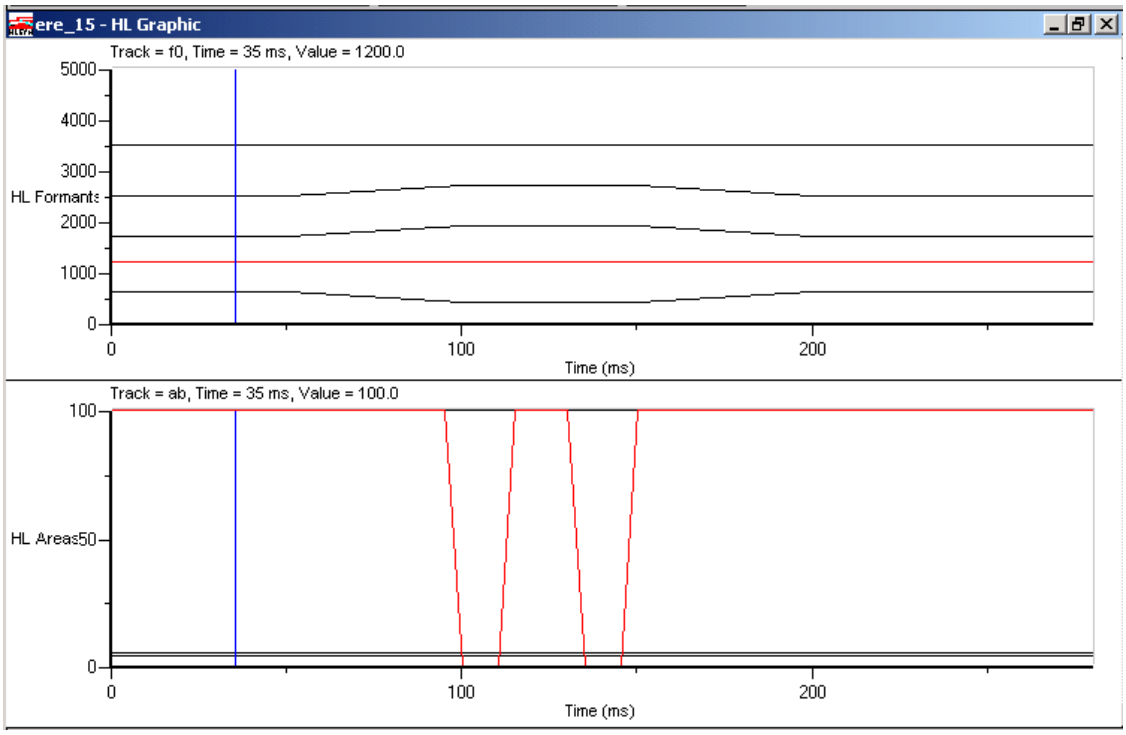
Vytvořili jsme i variantu se dvěma kmity. Podle reálného vzoru trvají oba závěry včetně fází otevírání a zavírání po 20ms. Rychlost kmitání je tak větší než u jednokmitného /r/. Stevens sice uvádí trvání vícekmitného /r/ a 200ms, ale v jeho případě jde o geminátu v italštině. Fáze zavírání a otevírání jsme tak nastavili na 5ms, fázi závěru na 10ms. Pseudovokalická fáze mezi kmity sestává z fáze zavírání prvního kmitu a fáze otevírání druhého kmitu a neprojeví se tak plně formantová struktura.



Graf 8 Parametry syntézy sekvence /ere/, dvojkmitná varianta (ere_14)

2.1.3.11 Pokus patnáctý - dvojkmitná varianta s delší mezikmitnou fází (ere_15)

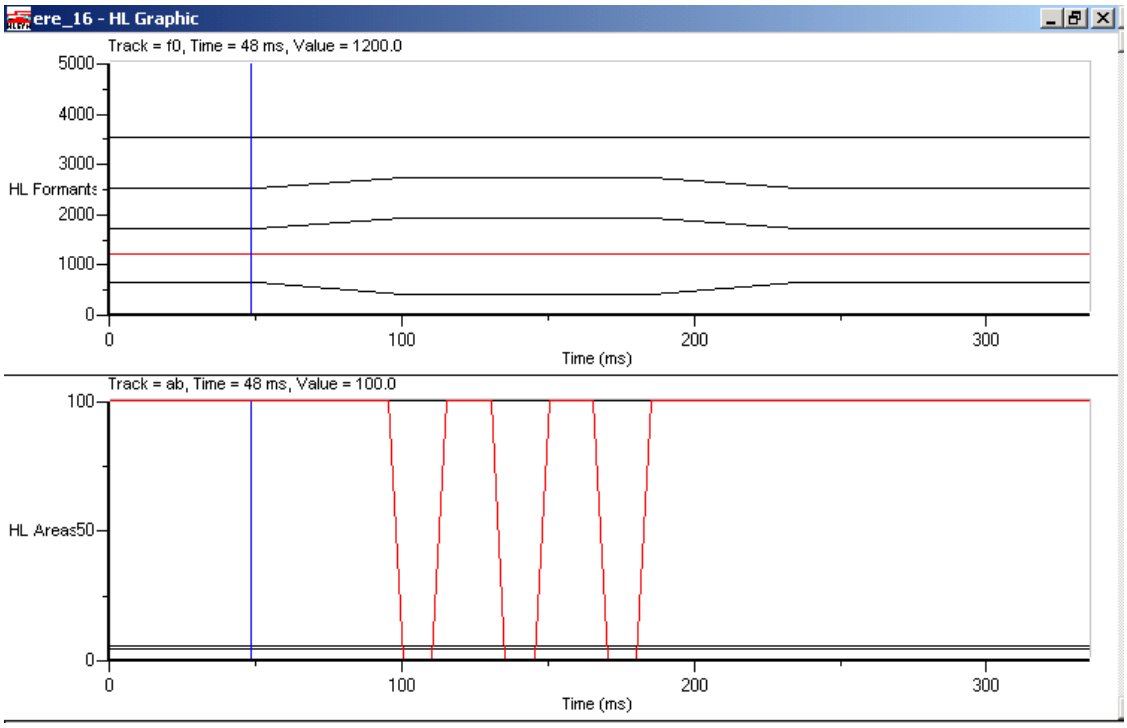
Mezikmitnou fázi prodloužili na 25ms, z čehož prvních a posledních 5ms zahrnuje otevírání, resp. zavírání.



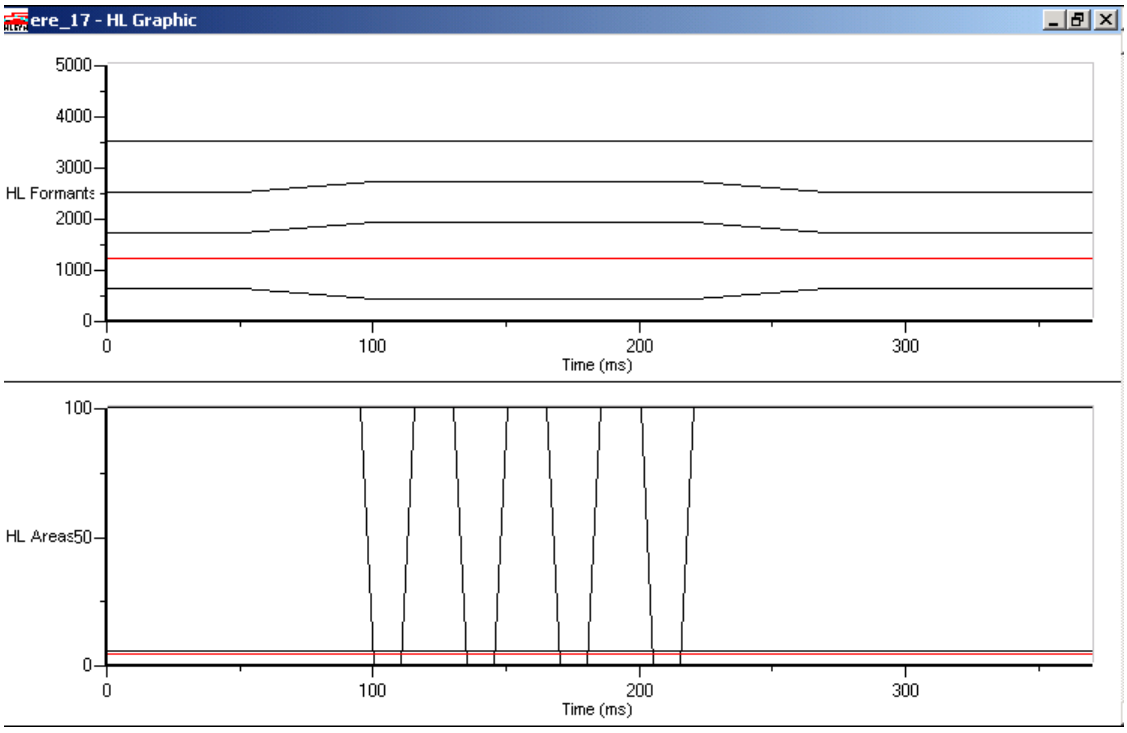
Graf 9 Parametry syntézy sekvence /ere/, dvojkmitná varianta s delší mezikmitnou fází (ere_15)

2.1.3.12 Pokus šestnáctý a sedmnáctý (ere_16 – ere_17)

Pro úplnost jsme vytvořili podle teorií Maláče a Borovičkové (1967) a Palkové (1994) variantu 4-kmitnou, respektive 3-kmitnou. Sedmikmitnou variantu podle Chlumského (1911) jsme již netvořili. Ponechali jsme trvání kmitů i délku mezikmitné fáze, výsledná hláska je tak delší.



Graf 10 Parametry syntézy sekvence /ere/, tříkmitná varianta (ere_16)



Graf 11 Parametry syntézy sekvence /ere/, čtyřkmitná varianta (ere_17)

Pro přehlednost uvádíme tabulku základních hodnot jednotlivých variant, barevně je vždy hodnota, která je pro danou variantu charakteristická.

poznámka	číslo	počet kmitů	f1	f2	f3	doba tranzientů	délka závěru	nástup závěru	mezivokalická f.	trvání
krátké /d/	1	1	400	1800	2700	50	25	25	0	75
krátké /d/ posunuté vzad	2	1	400	**1900**	2700	50	25	25	0	75
rychlejší zavírání/otevírání	3	1	400	1900	2700	50	25	**7**	0	40
kratší závěr	4	1	400	1900	2700	50	**10**	7	0	25
f2 1800	5	1	400	**1800**	2700	50	10	7	0	25
f2 2000	6	1	400	**2000**	2700	50	10	7	0	25
f2 2200	7	1	400	**2200**	2700	50	10	7	0	25
krátké /d/ formanty M&B	8	1	**600**	**1600**	**2500**	50	25	25	0	75
rychlejší zavírání/otevírání formanty M&B	9	1	**600**	**1600**	**2500**	50	25	**7**	0	40
kratší závěr formanty M&B	10	1	**600**	**1600**	**2500**	50	**10**	7	0	25
ere_5 změněn f1	11	1	**600**	1900	2700	50	10	7	0	25
ere_5 změněn f2	12	1	400	**1600**	2700	50	10	7	0	25
ere_5 změněn f3	13	1	400	1900	**2500**	50	10	7	0	25
dvojkmitné s defektní mezifází	14	**2**	400	1900	2700	50	10	5	**10**	40
dvojkmitné	15	**2**	400	1900	2700	50	10	5	**25**	55
Tříkmitné	16	**3**	400	1900	2700	50	10	5	25	70
čtyřkmitné	17	**4**	400	1900	2700	50	10	5	25	85

Tabulka 1 Přehled variant hlásky /r/

2.1.4 Syntéza /ř/

Foném /ř/ se v češtině vyskytuje ve dvou alofonech, a to znělé a neznělé variantě s komplementární distribucí. V syntéze tuto skutečnost odráží parametr ag.

2.1.4.1 Znělé /ř/

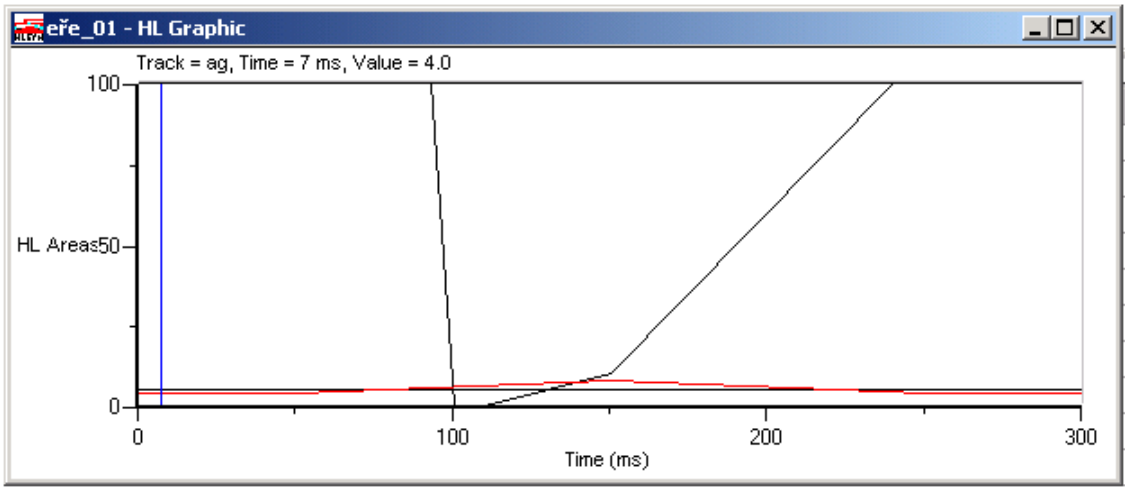
2.1.4.1.1 Pokus první /r+ž/, alveolární (eře_01)

Ačkoliv již Chlumský (1911) tvrdí, že hláska /ř/ se nedá považovat za afrikátu /rž/, zvolili jsme tuto variantu pro její jednoduchost a percepční blízkost za výchozí. Vycházeli jsme z jednokmitného /r/ (varianta ere_04), závěr jsme ponechali dlouhý 10ms, frikativní část odpovídající hlásce /ž/ jsme pak protáhli na 90ms, což je v rozporu s Chlumským, který tvrdí, že pouze frikativní část je kratší, ovšem více odpovídá reálnému spektrogramu. Naopak je to méně než uvádí Maláč a Borovičková (1967) (195ms). Celá hláska má tak 100ms (plus 7ms na začátku na fázi zavírání), což je v souladu alespoň s Chlumského tvrzením, které říká, že /ř/ je delší než /r/.

Tranzienty začínají 50ms před vibrantní částí a končí 50ms po šumové části.

Frikativy jsou tvořeny změnou parametru ag – ve své polovině mají znělé hlásky hodnotu 8, původní hodnota (4) je naposledy fixována 50ms před začátkem hlásky a 50ms po konci hlásky, hranice jsou dopočítány.

Hodnota ab se pohybuje od původní hodnoty v -40ms do maxima 10 uprostřed hlásky, a znovu původní hodnotu nabývá 40ms po skončení hlásky. V našem případě, kdy ab tvoří konstrikci vibrantní části, nejde samozřejmě dodržet čas nástupu, je tedy naposledy fixována hodnota 0 na konci závěru.



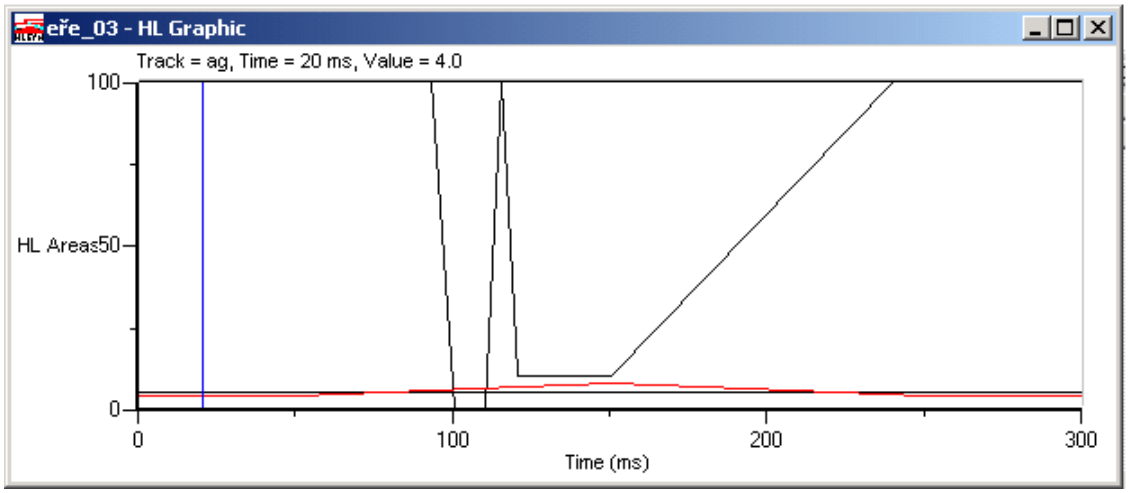
Graf 12 Parametry syntézy sekvence /eře/

2.1.4.1.2 Pokus druhý – post-alveolární místo tvoření (eře_02)

Druhý formant jsme v celém trvání nastavili na hodnotu odpovídající postalveolárního místa tvoření (2000Hz) (což odpovídá poznatkům Hájkové, 2010).

2.1.4.1.3 Pokus třetí – dokončení cyklu vibrace (eře_03)

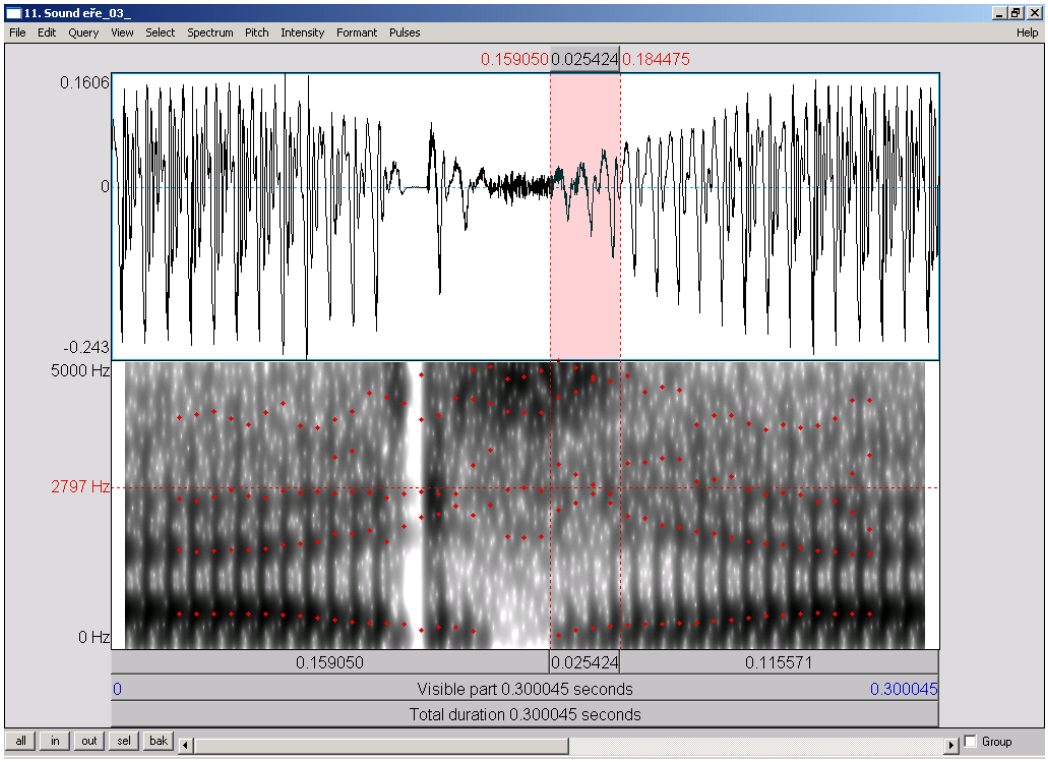
Protože je potřeba posílit vibrantní část, další pokus spočívá v dokončení cyklu vibrace, čili přechodu z hodnoty 0 u parametru ab na 100 a poté teprve snížení na 10, obojí po 5ms.

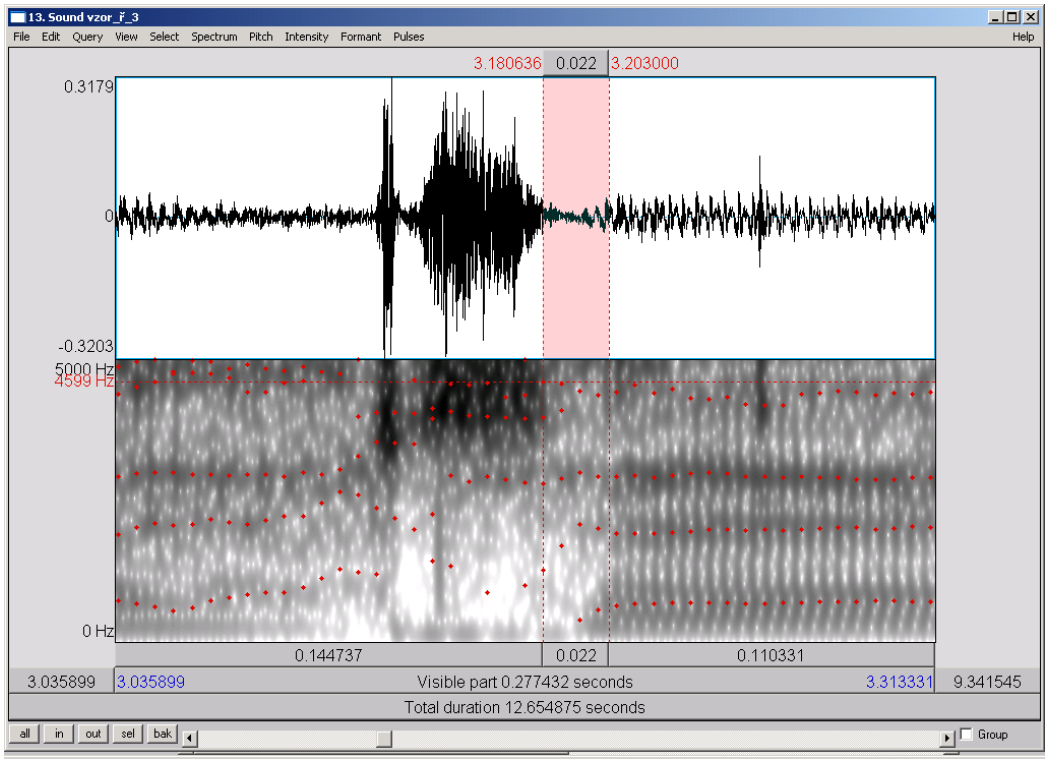


Graf 13 Parametry syntézy sekvence /eře/, dokončení cyklu vibrace

2.1.4.1.4 Pokus čtvrtý – bez překryvu šumu a nástupu formantů (eře_04)

Ze srovnání s reálným spektrogramem (obr. 22 a 23) je patrné, že v reálu nedochází k takovému překryvu šumu a nástupu formantů jako u syntetizované varianty.

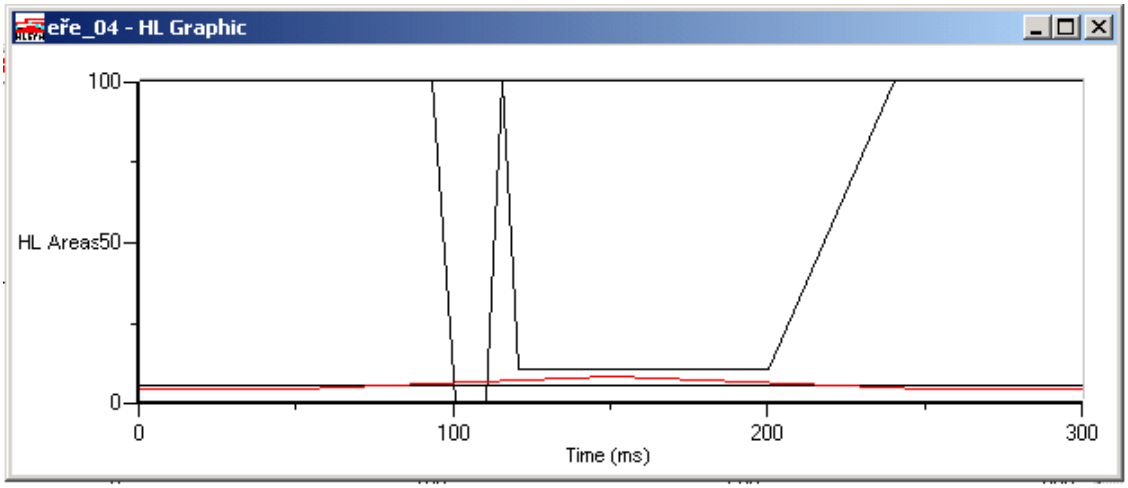




Obr. 22 a 23 – srovnání eře_03 a reálného vzoru, v syntetizované variantě je zvýrazněna část, kde se překrývá šum a již nastupuje formantová struktura následující

hlásky, v reálném vzoru je zvýrazněna naopak část, kde již šum ustoupil, ale formantová struktura samohlásky je ještě neúplná

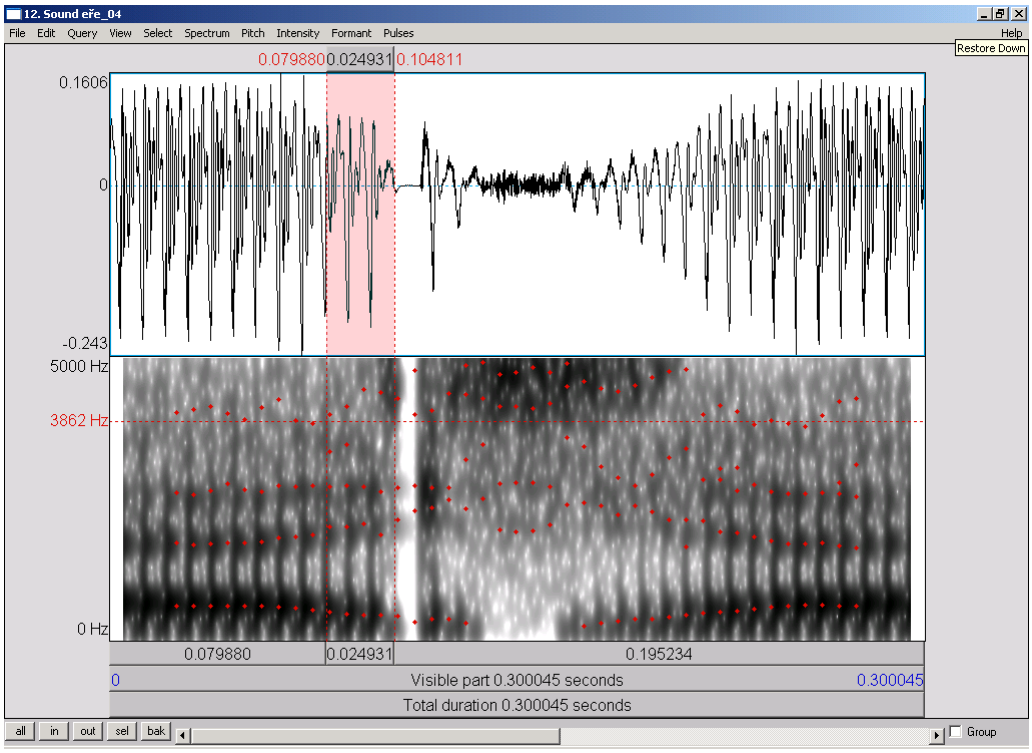
Nechali jsme tedy hodnotu 10 u parametru ab až do konce šumové části hlásky, přechod na hodnotu 100 je tak mnohem rychlejší (během 40ms místo 90ms).

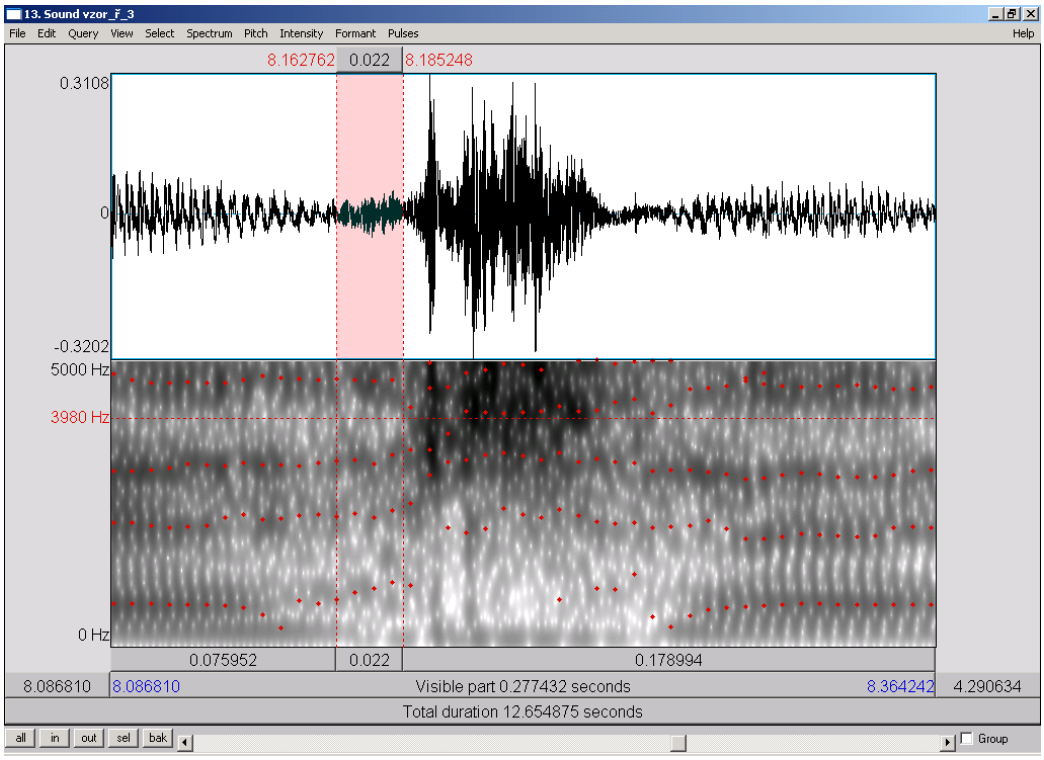


Graf 14 Parametry syntézy sekvence /eře/, bez překryvu šumu a nástupu formantů

2.1.4.1.5 Pokus pátý – předcházející šum (eře_05)

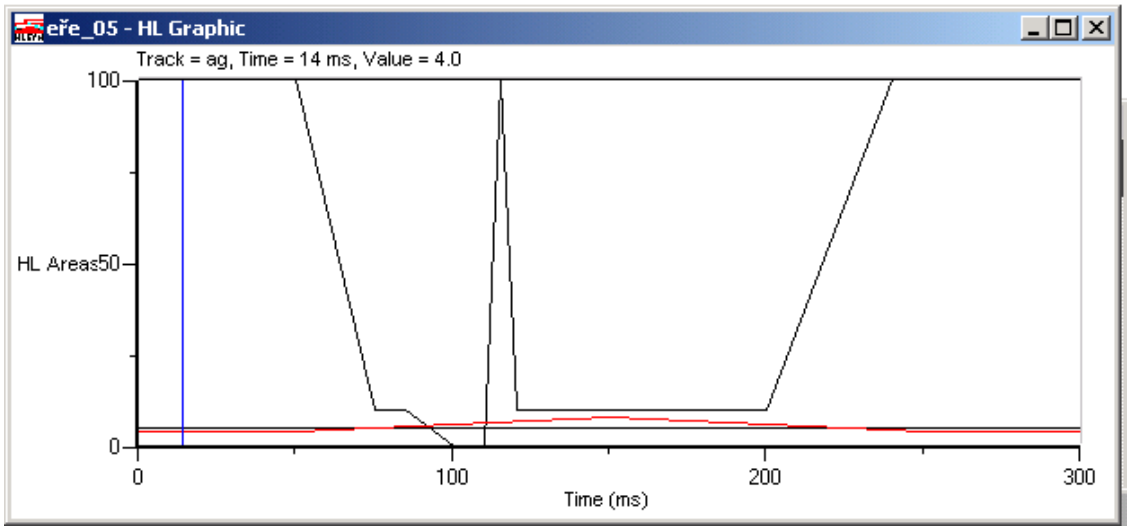
Z dalšího srovnání s reálným spektrogramem (obr. 24 a 25) je vidět, že zatímco v syntetizované variantě přechází plná formantová struktura předchozího /e/ přímo do vibrantní části, v reálné variant předchází vibraci ještě zhruba 20ms šum.





Obr. 24 a 25 – první je syntetizovaná varianta, zvýrazněna je část, kde plná formantová struktura přechází přímo do závěru, zatímco v druhém, reálném spektrogramu je vidět, že zhruba 20ms před závěrem se formantová struktura začíná vytrácet a percepčně je zde již šum

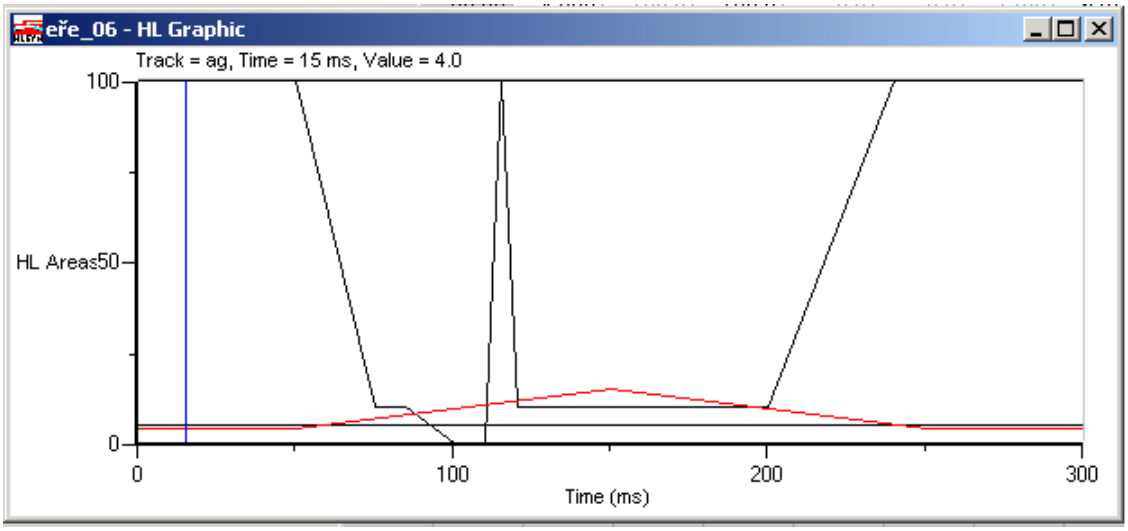
Hodnotu parametru ab jsme tak stáhli na 10 už 25 ms před začátkem vibrace a nechali z hodnoty 10 přecházet během 15ms na 0.



Graf 15 Parametry syntézy sekvence /eře/, předcházející šum

2.1.4.1.6 Pokus šestý – míra ag (eře_06)

Ze sledování více provedení /ř/ v češtině jsme dospěli k názoru, že i znělá varianta je méně znělá než ostatní znělé frikativy, často je i intervokalické /ř/ z důvodu artikulační náročnosti desonorizováno. Toto pozorování jsme se rozhodli reflektovat zvýšením hodnoty parametru ag na hodnotu 15 (místo 8, která platí pro znělé frikativy).



Graf 16 Parametry syntézy sekvence /eře/, míra ag

2.1.4.1.7 Pokus sedmý – formanty podle Maláče a Borovičkové (eře_07)

Borovičková a Maláč udávají ve své zprávě formanty znělého /ř/ v hodnotách 600, 1600 a 2000Hz. První a druhý formant jsme nastavili podle jejich hodnot, ovšem třetí jsme nechali na hodnotě 2700, jelikož při hodnotě 2000 je zvuk zcela hlásce /ř/ nepodobný.

2.1.4.1.8 Pokus osmý a devátý – srovnání jednotlivých formantů (eře_08 – eře_09)

V osmém pokusu jsme nastavili pouze f1 na hodnotu podle B&M, v devátém f2. Variantu pro změněné pouze f3 jsme nedělali, jelikož by mělo stejnou hodnotu jako f2.

2.1.4.2 Neznělé /ř/

Pokusy jsou ve všech parametrech kromě ag identické se znělým /ř/, pouze hodnota ag směřuje místo do 8, resp. 15 do hodnoty 20. Pokus šest je při změně ag identický s pokusem pět, proto v neznělé variantě není.

2.2 Ověření percepčním testem

Navržené a syntetizované varianty jsme ověřili percepčním testem, který se zaměřil na percepční přirozenost syntetizovaných zvuků. Pro co nejjemnější stanovení škály přirozenosti jsme zvolili test porovnávání páru (category comparision rating, CCR), posluchači vždy slyšeli dvě stejná slova s odlišnou cílovou hláskou a měli zvolit, zda je přirozenější varianta A či B. Možnost „stejně přirozené“ jsme jim nenabízeli, protože rozdíly jsou často skutečně malé a zbytečně by je možnost sváděla k této odpovědi. Předpokládali jsme totiž, že pokud jsou nějaké dvě varianty stejně přirozené a rozdíl mezi nimi je mizivý, bude zaprvé jejich index přirozenosti (viz níže) na škále podobný, a zadruhé, jejich srovnání dopadne podle principu náhody 50:50.

2.2.1 Sestavení testu

Původní záměr by v testu konfrontovat každou variantu se všemi ostatními, ovšem z důvodu většího množství variant u hlásky /r/ by test narostl netestovatelné velikosti, zvolili jsme tedy alternativní rozdělení cílových hlásek.

V testu A byly vzájemně porovnávány varianty 1-10, což jsou v podstatě všechny zásadní jednokmitné varianty.
V testu B byla porovnávána varianta 4 s variantami 11-13, což jsou varianty, kde je měněn jeden z parametrů ve prospěch hodnot uváděných Maláčem a Borovičkovou (1967).
V testu C byla porovnávána varianta 4 s variantami 14-17, tedy jednokmitná vs. vícekmitné varianty.

2.2.2 Přepočet na index přirozenosti

Pro účel vyhodnocení výsledků percepčního testu a srovnání jednotlivých variant jsme zavedli pojem „index přirozenosti“. Pro každou variantu se tato hodnota dostane následujícím způsobem: sečte se počet případů, kdy hláska zněla přirozeněji než její konkurent.
Varianty se pak uspořádají podle této hodnoty, čímž dostaneme škálu přirozenosti.

2.2.3 Provedení percepčního testu

Test byl zadán online. Na internetové stránky byly umístěny sobory A-E a jeden zácvičný společně s pokyny a zaškrtávacími políčky. Po vyplnění testu byly výsledky odeslány do tabulky pomocí Google docs formuláře.
Testu se zúčastnilo celkem 20 osob ve věku 16-56 let, věkový průměr byl 30let. Osoby byly původem z různých míst republiky. Test vyplnily převážně ženy (16 žen, 4 muži). Nemyslíme si však, že by některá z těchto kategorií měla vliv na hodnocení hlásek a ani výsledky nic takového nenaznačují.

3 Výsledky

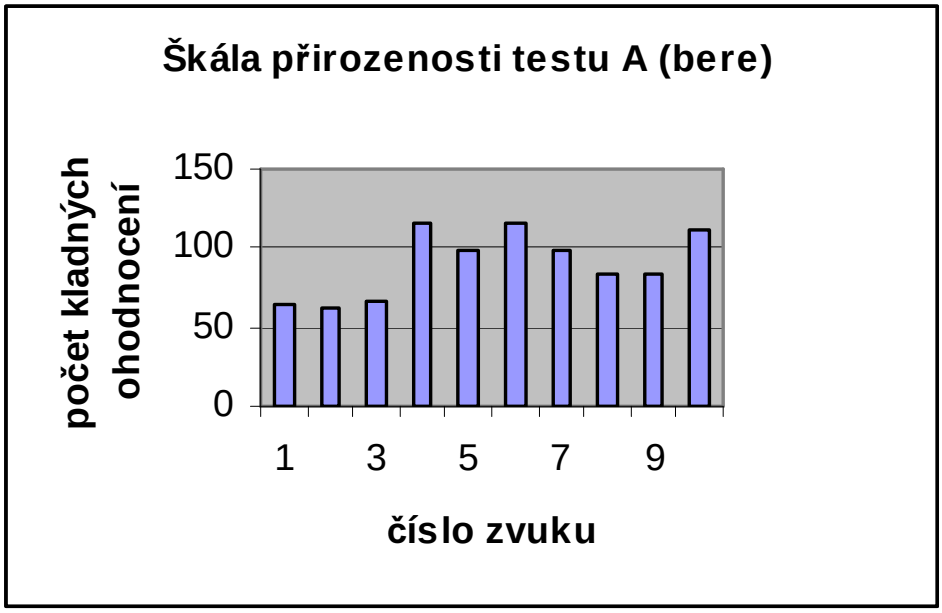
Kompletní odpovědi a výsledky jsou v elektronické tabulce na CD. Signifikantnost jsme testovali testem chí-kvadrát.

3.1 Test A – základní testování variant hlásky /r/

V testu A jsme dostali následující výsledky:

číslo varianty	**popis varianty**	**počet poz. hodnocení**
1	krátké /d/	64
2	krátké /d/ posunuté vzad	63
3	rychlejší zavírání/otevírání	67
4	kratší závěr	116
5	f2 1800	99
6	f2 2000	115
7	f2 2200	98
8	krátké /d/ formanty M&B	83
9	rychlejší zavírání/otevírání formanty M&B	83
10	kratší závěr formanty M&B	112

Tabulka 2: Výsledky testu A



Graf 17: Škála přirozenosti testu A

Jak je vidět z výsledků, první dvě úpravy (posunutí vzad a rychlejší otevírání přirozenost hlásky nijak významně nezvýšily. Oproti tomu třetí úprava, zkrácení závěru, byla podstatná.

Ze srovnání tří hodnot druhého formantu (1800, 2000, 2200Hz – varianty 5, 6, 7, a 1900Hz – varianta 4) vidíme, že hodnoty 1900Hz a 2000Hz vycházejí lépe než hodnoty 1800 a 2200Hz. Je zajímavé, že je zde patrnější rozdíl mezi hodnotou 1800Hz a 1900Hz, ačkoliv mezi prvními dvěma variantami, které byly ještě blízko /d/, takový rozdíl vnímán nebyl. Patrně je tomu tak proto, že byla varianta rovnou zavržena jako hlásce /r/ nepodobná a již posluchači neřešili místo artikulace.

Srovnání variant 1-8, 3-9 a 4-10 (naše hodnoty formantů vs. hodnoty Maláče a Borovičkové) neukazuje žádné výrazné tendence pro jednu či druhou variantu. U nedokonalých variant 1 a 3, respektive 8 a 9, je nastavení formantů podle Maláče a Borovičkové hodnoceno lépe, ovšem důležitější je srovnání variant 4 a 10, které posluchači hodnotili jako více r-ové. Zde vycházejí výsledky téměř nastejno, s velmi mírnou převahou pro variantu 4. Když se podíváme na výsledky srovnání jednotlivých párů, zjistíme, že „duel“ 4-10 dopadl poměrem 13:7, tedy zde má varianta 4 větší převahu.

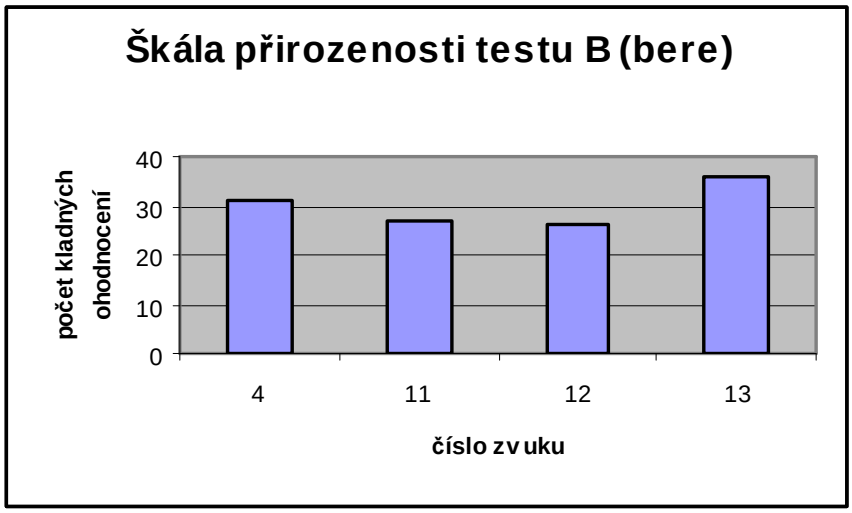
Výsledky testu A mají p = 0,0000013, výsledky tedy můžeme považovat za vysoce signifikantní.

3.2 Test B – testování jednotlivých formantů hlásky /r/

Test B přinesl následující výsledky:

číslo varianty	popis varianty	počet poz. hodnocení
4	kratší závěr	31
11	ere_4 změněn f1	27
12	ere_4 změněn f2	26
13	ere_4 změněn f3	36

Tabulka 3: Výsledky testu B



Graf 18: Škála přirozenosti testu B

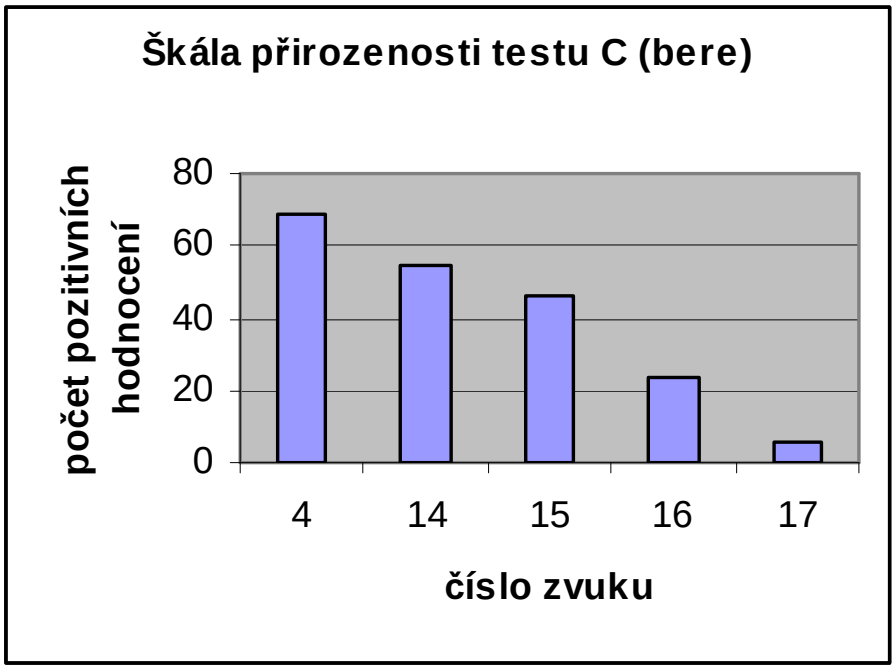
Výsledky testovaných variant vyšly velmi podobně, je vidět nepatrná tendence hodnotit změněný první a druhý formant negativněji, a třetí formant pozitivněji.
Pro tuto tabulku vychází p = 0,009, což je stále signifikantní výsledek.

3.3 Test C – testování vícekmitných variant hlásky /r/

Test srovnání různého počtu kmitů dopadl následovně:

číslo varianty	popis varianty	počet poz. hodnocení
4	kratší závěr	69
14	dvojkmitné s defektní mezifází	55
15	dvojkmitné	46
16	tříkmitné	24
17	čtyřkmitné	6

Tabulka 4: Výsledky testu C



Graf 19 Škála přirozenosti testu C

Výsledky testu hovoří celkem jasně – čím více kmitů (v případě varianty 14 čím delší mezifáze), tím méně přirozené. Je samozřejmě možné, že se tyto výsledky nevztahují ani tak k počtu kmitů, jako celkovému trvání hlásky. Vícekmitné varianty jsou totiž delší, protože časy kmitů zůstaly zachovány, takže se nabízí i interpretace, že čím je /r/ delší (od nějaké hraniční hodnoty), tím je méně přirozené. Pro rozhodnutí této dvojznačnosti by bylo potřeba nesyntetizovat a percepčně otestovat vícekmitné varianty, které by zachovávaly trvání jednokmitné varianty a jejich kmity by tak byly rychlejší.

Signifikantnost vychází p = 2,5E-12, což značí opravdu velmi malou pravděpodobnost náhody.

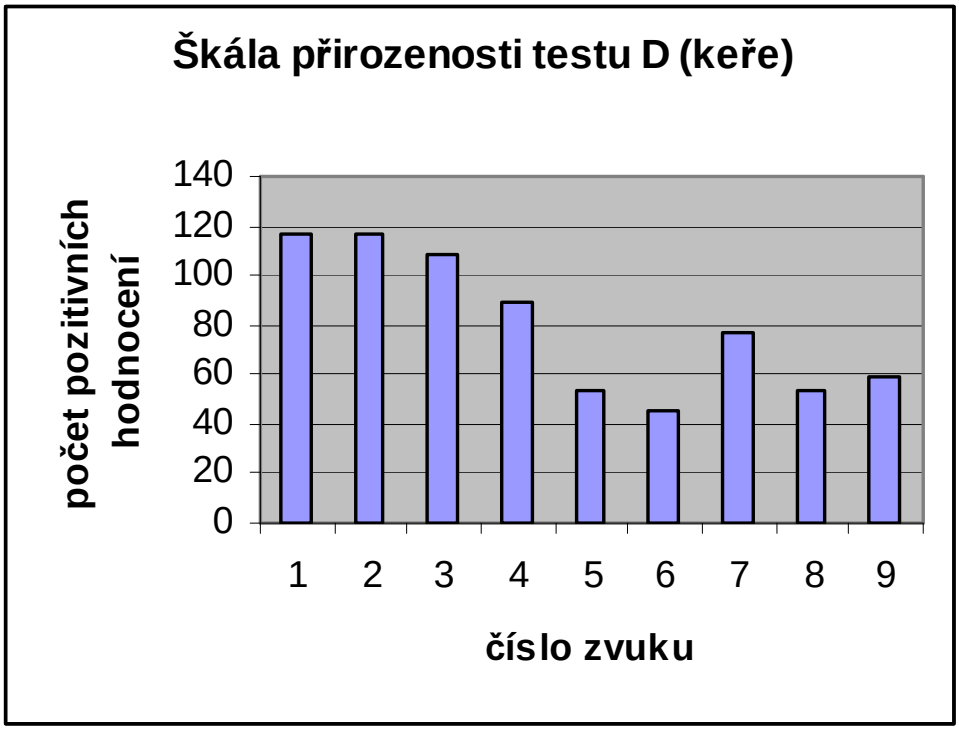
3.4 Test D – testování znělého /ř/

Percepční test hlásky /ř/ dopadl překvapivě:

číslo varianty	popis varianty	počet poz. hodnocení
1	alveolární /r+ž/	117
2	postalveolární /ř+ž/	117
3	dokončení cyklu vibrace	109

4	bez překryvu šumu a nástupu formantů	89
5	předcházející šum	53
6	snížení znělosti	45
7	f1 a f2 podle B&M	77
8	f1 podle B&M	54
9	f2 podle B&M	59

Tab. 5 Výsledky testu D



Graf 20 Škála přirozenosti testu D:

Ačkoliv jsme očekávali, že podobně jako u hlásky /r/, budou vyšší, zdokonalované varianty, vnímány jako přirozenější, percepční test ukázal pravý opak. Mezi první a druhou variantou, lišící se v místě artikulace (hodnota druhého formantu) není žádný rozdíl, i porovnávání páru 1 vs. 2 dopadlo téměř nerozhodně (11:9) – je otázka, zda jsou posluchači u hlásky /ř/ ohledně místa artikulace benevolentnější, nebo byla chyba v syntéze (rozdíl v druhém formantu měl být patrnější, případně je místo artikulace v tomto případě silněji závislé na některém z dalších parametrů). S každou další úpravou až do varianty 6 se pak vnímání přirozenosti snižuje. Kupodivu změna formantů na hodnoty podle Maláče a Borovičkové (1967) přirozenost zvýší. Z testů varianty 8 a 9 je možné odhadovat, že větší podíl na tom má změna druhého formantu.

Podle studie Hájkové (2010) by mělo být /ř/ artikulováno spíše vzaději, je tak vnímání varianty /ř/ s f2 = 1600Hz jako přirozenější velmi překvapivé a nemáme pro něj žádné teoretické vysvětlení.

Ke klesající přirozenosti pokusů máme pouze velmi dohadové vysvětlení. Z několika náhodných hovorů s respondenty po testu jsme zaznamenali opakující se komentáře typu „nevím, jestli jsem to /ř/ vyplnil/a správně, protože jsem zaškrtával/a to, co mi znělo sice míň jako /ř/, ale znělo to přirozeněji“. Je tedy možné, že není zásadní chyba v postupu úprav, ovšem úpravy nejsou dokonalé a dávají zvuku příliš umělý charakter, který posluchače odrazuje.

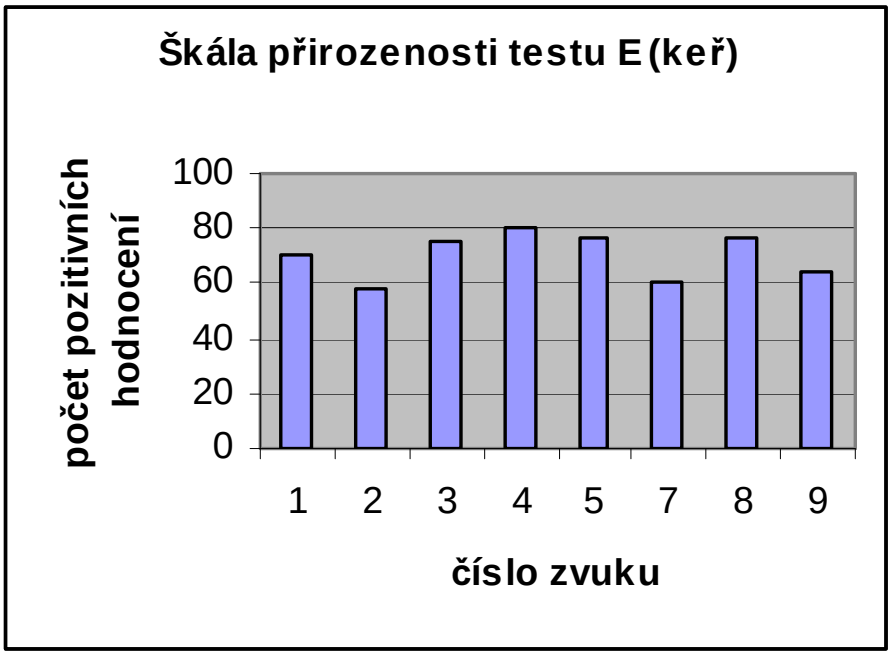
Výsledky testu vycházejí s p = 6,79E-15, což opět vylučuje náhodu.

3.5 Test E – testování neznělého /ř/

Výsledky dopadly takto:

číslo varianty	**popis varianty**	**počet poz. hodnocení**
1	alveolární /r+ž/	70
2	postalveolární /ř+ž/	58
3	dokončení cyklu vibrace	75
4	bez překryvu šumu a nástupu formantů	80
5	předcházející šum	76
7	f1 a f2 podle B&M	61
8	f1 podle B&M	76
9	f2 podle B&M	64

Tab. 6 – Výsledky testu E



Graf 21: Škála přirozenosti testu E

U neznělé varianty hlásky /ř/ se stejným postupem „vylepšování“ bychom očekávali podobné výsledky, ale překvapivě zde sestupná tendence patrná není. Zajímavé je, že zatímco u znělé varianty formanty podle Maláče a Borovičkové (1967) přirozenost mírně zvýšily, u neznělé varianty ji naopak snižují, a to hodnota druhého formantu výrazněji.

Lze se tedy domýšlet, že pro neznělé hlásky nejsou rozdíly, které jsme ve variantách zaváděli, natolik patrné, aby je posluchači spolehlivě odlišili. Alternativní vysvětlení by bylo, že za toto smývání rozdílů může finální pozice ve slově. Třetím vysvětlením může být zařazení na konec testu a únava posluchačů.

U tohoto testu vychází p = 0,47, což znamená, že rozdíly jsou natolik minimální, že z nich nelze dělat žádné závěry.

4 Diskuse

Je třeba předně říci, že tato práce není technicky vyčerpávající. HL systém je značně komplikovaný a detailní porozumění jeho parametrům by přesahovalo rámec bakalářské práce. Navíc k programu neexistuje kompletní dokumentace, materiál dostupný v manuálu je kusý a neúplný. S parametry je tedy nakládáno spíše intuitivním způsobem a metodou zkoušení různých variant, než přímým aplikováním teorie.

Dále je třeba podotknout, že práce byla zaměřena na vyzkoušení možností syntézy pro české vibranty, nebyla tedy primárně určena k ověření detailních rozdílů mezi jednotlivými variantami a zhodnocení, jaké hodnoty všech parametrů čeští mluvčí vnímají jako nejpřirozenější. Takový výzkum by byl nepochybně zajímavý a přínosný, ale bylo by jej třeba realizovat ve více etapách percepčních testů, vždy zaměřených pouze na jeden parametr. Pokud bychom měli testovat vše najednou, byl by percepční test pro posluchače neúnosně veliký a výsledná data zkreslená.

5 Závěr

V naší práci jsme se pokusili na základě obecných znalostí o českých vibrantách a studia několika jejich konkrétních realizací stanovit parametry pro HL syntézu, která spojuje čistě formantovou a artikulační syntézu zavedením HL parametrů odrážejících artikulační nastavení, které jsou přepočítávány na parametry klasické formantové klattovské syntézy.

Pro každou hlásku (/r/, znělé /ř/, neznělé /ř/) jsme syntetizovali postupně několik variant, jejichž vydařenost jsme pak otestovali percepčním testem. Výsledky testu měly ukázat přirozenost jednotlivých cílových hlásek.

Pro hlásku /r/ dopadl percepční test podle očekávání – zlepšení (zkrácení doby otevírání a zavírání, kratší doba závěru) byla hodnocena zvýšením indexu přirozenosti. Testováním několika hodnot druhého formantu, který koresponduje s místem tvoření, jsme zjistili, že nejlépe jsou vnímány hodnoty 1900-2000Hz. Srovnáním s vícekmitnými variantami jsme dospěli k zjištění, že posluchači vnímají nejpřirozeněji jednokmitné /r/, a čím více kmitů, tím horší hodnocení.

Pro znělé /ř/ ovšem percepční test ukázal překvapivě zcela opačné výsledky – čím více vylepšení, tím méně přirozené. Tento výsledek si zdůvodňujeme tím, že vylepšené varianty sice nejspíš odrážely vlastnosti reálné hlásky, ale ne dokonale, což posluchače přimělo k horšímu hodnocení než u hlásek, které sice zněly méně jako /ř/, ale byly více přirozené.

Pro neznělé /ř/ byly všechny varianty hodnoceny s velmi malými rozdíly přirozenosti. Toto setření rozdílů lze přikládat buď finální pozici ve slově nebo absenci znělosti.

Seznam literatury

- Borovičková, B., Maláč, V. (1967) The Spectral Analysis of Czech Sound Combination. In: Rozpravy Československé akademie věd, ročník 77, sešit 14. Academia.
- Fant, G. (1960) Acoustic Theory of Speech Production. Mouton de Gruyter, Gravenhage
- Hanzlíček, Z. (2010) Czech HMM-Based Speech Synthesis. In: Sojka, P. (Ed.) (2010) TSD 2010, LNAI 6231, str. 291-298
- Chlumský, J. (1911). Une varieté peu conuue de l'R linguale. Revue de phonétique, 1, str. 33-67.
- Isačenko, A. V. (1965) Zur Akustik des tschechischen ř-Lautes. In: Bethge, W., Mal,berg, B., Pilch, H., Zwirner E. (eds.), Phonetika 12, str. 1-12.
- Kučera, H. (1961). The Phonology of Czech. S-Gravenhage : Mouton & Co., str. 30.
- Ladefoged, P. & Maddieson, I. (1996). The sounds of the world's languages. Oxford: Blackwell.
- Lindau, M. (1980). The story of /r/. The Journal of the Acoustical Society of America 67, S27
- Machač, P. (2009) Implications of Acoustic Variation for the Segmentation of the Czech Trill /r/. In: Esposito, A., Vích, R. (Eds.) Cross-Modal Analysis, LNAI 5641, str. 173-181
- McGowan, R. S. (1992). Tongue-tip trills and vocal-tract wall compliance. Acustical Society of America, Vol. 91, No. 5, str. 2903-2910
- Palková, Z. (1994) Fonetika a fonologie češtiny. Karolinum
- Psutka, J a kol. (2006). Mluvíme s počítačem česky, Academia
- Ptáček, M. (1996) Czech Speech Synthesis. In: Palková, Z. (ed.) Phonetica Pragensia X, 1, Acta Universitatis Carolinae, str. 231-244
- Recasens, D. (1999) The study of /ɾ/ and /r/ in the light of the "DAC" coarticulation model, Journal of Phonetics, 1999, 27, str. 143-169
- Sensimetrics (2004). HLsyn, High-Level Speech Synthesizer User Interface Manual. Somerville, MA.
- Stevens, K. & Bickley, C. (1991). Constraints among parameters simplify control of Klatt formant synthesizer. Journal of Phonetics, 19, str. 161-174.
- Trávníček, F. (1932) Úvod do české fonetiky. Praha: Česká graická unie, str. 52-53
- Westbury, J. R. (1983). Enlargement of supraglottal cavity and its relation to stop consonant voicing. Acustical Society of America, Vol. 73, No. 4, str. 1322-1336